Developing Materials Informatics Workbench for Expediting the Discovery of Novel Compound Materials

Kwok Wai Steny Cheung

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in September 2009

The Australian Institute for Bioengineering and Nanotechnology

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the *Copyright Act 1968*.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material.

Statement of Contributions to Jointly Authored Works Contained in the Thesis

No jointly-authored works

Statement of Contributions by Others to the Thesis as a Whole

No contributions by others.

Statement of Parts of the Thesis Submitted to Qualify for the Award of Another Degree

None.

Published Works by the Author Incorporated into the Thesis

J. Hunter and K. Cheung. 'Generating eScience Workflows from Statistical Analysis of Prior Data' APAC'05. Royal Pines Resort, Gold Coast. 26–29 Sept 2005. ― Partially incorporated as paragraphs in Chapter 2.

J. Hunter, K. Cheung, S. Little, and J. Drennan. 'FUSION – A Knowledge Management System for Fuel Cell Optimization' International Conference on Solid State Ionics. Applications: Fuel Cells. Baden-Baden. July 2005, p. 544. ― Partially incorporated as paragraphs in Chapter 2.

K. Cheung, J. Drennan, J. Hunter. 'Towards an Ontology for Data-driven Discovery of New Materials', Semantic Scientific Knowledge Integration, AAAI/SSS Workshop, Stanford University, 26–28 Mar 2008. ― Incorporated as Chapter 3.

K. Cheung, J. Hunter, J. Drennan. 'MatSeek – An Ontology-based Federated Search Interface for Materials Scientists', IEEE Intelligent Systems, Volume 24, Issue 1, pp. 47–56, January/February 2009. doi:10.1109/MIS.2009.13. ― Incorporated as Chapter 4.

K. Cheung, J. Hunter, J. Drennan. 'MatSeek – An Ontology-based Federated Search Interface for Materials Scientists', 21st International CODATA Conference, Ukraine, Kyiv, 5–8 Oct 2008. ― Incorporated as Chapter 4.

J. Hunter and K. Cheung, 'Provenance Explorer –A Graphical Interface for Constructing Scientific Publication Packages from Provenance Trails', Special Issue –Digital Libraries and eScience, International Journal on Digital Libraries, Volume 7, Numbers 1–2, October 2007. doi:10.1007/s00799-007-0018-5. ― Partially incorporated as paragraphs in Chapter 5.

K. Cheung and J. Hunter. 'Provenance Explorer –Customized Provenance Views Using Semantic Inferencing', Proceedings of the Fifth International Semantic Web Conference, ISWC2006. Athens, GA, USA. November 2006. pp. 215–227. doi:10.1007/11926078_16. ― Partially incorporated as paragraphs in Chapter 5.

K. Cheung and J. Hunter. 'Provenance Explorer –A Tool for Viewing Provenance Trails and Constructing Scientific Publication Packages'. Digital Library Goes e-Science (DLSci06).Alicante, Spain. September 2006,pp. 13–26. ― Partially incorporated as paragraphs in Chapter 5.

K. Cheung, J. Hunter, A. Lashtabeg, J. Drennan. 'SCOPE: A Scientific Compound Object Publishing and Editing System', International Journal of Digital Curation, vol. 3, no. 2, 2008. ― Incorporated as Chapter 5.

K. Cheung, J. Hunter, A. Lashtabeg, J. Drennan. 'SCOPE –A Scientific Compound Object Publishing and Editing System', 3rd International Digital Curation Conference, Washington DC, 11–13 Dec 2007. ― Incorporated as Chapter 5.

Additional Published Works by the Author Relevant to the Thesis but not Forming Part of it

None.

Acknowledgements

I would like to thank my supervisors, Prof John Drennan, Prof Jane Hunter and Dr Guido Governatori, for providing me with this opportunity and I gratefully acknowledge their contribution.

Especially, I would like to express my deepest gratitude to Prof Drennan's effort and time to steer me through a variety of challenges during my candidature.

I show my gratitude to Prof Jane Hunter for her input to and feedback on my research.

I also highly appreciate Dr Governatori's prompt and valuable input and feedback at the stage of the thesis writing.

At each stage of my research, the many discussions I had about research with Prof Robert Colomb were of great assistance to me as a new academic researcher.

This research was made possible through the financial support from the Australian Commonwealth Government (Department of Education, Employment and Workplace Relations) through the Australian Postgraduate Awards APA programme and the Australian Institute for Bioengineering and Nanotechnology at the University of Queensland.

I have to thank my wonderful wife and loving daughter — Christina and Sophia. Without their patience, understanding and support, I would not have completed this candidature.

I would like to dedicate this thesis to my mother and my father-in-law, who I know have helped me to finish this thesis despite them not being physically here.

This thesis has been edited by Tony Roberts.

Abstract

This project presents a Materials Informatics Workbench that resolves the challenges confronting materials scientists in the aspects of materials science data assimilation and dissemination. It adopts an approach that has ingeniously combined and extended the technologies of the Semantic Web, Web Service Business Process Execution Language (WSBPEL) and Open Archive Initiative Object Reuse and Exchange (OAI-ORE). These technologies enable the development of novel user interfaces and innovative algorithms and techniques behind the major components of the proposed workbench.

In recent years, materials scientists have been struggling with the challenge of dealing with the ever-increasing amount of complex materials science data that are available from online sources and generated by the high-throughput laboratory instruments and data-intensive software tools, respectively. Meanwhile, the funding organizations have encouraged, and even mandated, the sponsored researchers across many domains to make the scientifically-valuable data, together with the traditional scholarly publications, available to the public. This open access requirement provides the opportunity for materials scientists who are able to exploit the available data to expedite the discovery of novel compound materials. However, it also poses challenges for them. The materials scientists raise concerns about the difficulties of precisely locating and processing diverse, but related, data from different data sources and of effectively managing laboratory information and data. In addition, they also lack the simple tools for data access and publication, and require measures for Intellectual Property protection and standards for data sharing, exchange and reuse. The following paragraphs describe how the major workbench components resolve these challenges.

First, the materials science ontology, represented in the Web Ontology Language (OWL), enables, (1) the mapping between and the integration of the disparate materials science databases, (2) the modelling of experimental provenance information acquired in the physical and digital domains and, (3) the inferencing and extraction of new knowledge within the materials science domain. Next, the federated search interface based on the materials science ontology enables the materials scientists to search, retrieve, correlate and integrate diverse, but related, materials science data and information across disparate databases. Then, a workflow management system underpinning the WSBPEL engine is not only able to manage the scientific investigation process that incorporates multidisciplinary scientists distributed over a wide geographic region and self-contained computational services, but also systematically acquire the experimental data and information generated by the process.

Finally, the provenance-aware scientific compound-object publishing system provides the scientists with a view of the highly complex scientific workflow at multiple-grained levels. Thus, they can easily comprehend the science of the workflow, access experimental information and keep the confidential information from unauthorised viewers. It also enables the scientists to quickly and easily author and publish a scientific compound object that, (1) incorporates not only the internal experimental data with the provenance information from the rendered view of a scientific experimental workflow, but also external digital objects with the metadata, for example, published scholarly papers discoverable via the World Wide Web (the Web), (2) is self- contained and explanatory with IP protection and, (3) is guaranteed to be disseminated widely on the Web.

The prototype systems of the major workbench components have been developed. The quality of the material science ontology has been assessed, based on Gruber's principles for the design of ontologies used for knowledge–sharing, while its applicability has been evaluated through two of the workbench components, the ontology-based federated search interface and the provenance-aware scientific compound object publishing system.

Those prototype systems have been deployed within a team of fuel cell scientists working within the Australian Institute for Bioengineering and Nanotechnology (AIBN) at the University of Queensland. Following the user evaluation, the overall feedback to date has been very positive. First, the scientists were impressed with the convenience of the ontology-based federated search interface because of the easy and quick access to the integrated databases and analytical tools. Next, they felt the surge of the relief that the complex compound synthesis process could be managed by and monitored through the WSBPEL workflow management system. They were also excited because the system is able to systematically acquire huge amounts of complex experimental data produced by self-contained computational services that is no longer handled manually with paper-based laboratory notebooks. Finally, the scientific compound object publishing system inspired them to publish their data voluntarily, because it provides them with a scientific-friendly and intuitive interface that enables scientists to, (1) intuitively access experimental data and information, (2) author self-contained and explanatory scientific compound objects that incorporate experimental data and information about research outcomes, and published scholarly papers and peer-reviewed datasets to strengthen those outcomes, (3) enforce proper measures for IP protection, (4) comply those objects with the Open Archives Initiative Protocol – Object Exchange and Reuse (OAI-ORE) to

maximize its dissemination over the Web and,(5) ingest those objects into a Fedora-based digital library.

Keywords

semantic web, web services, wsbpel, materials informatics, oai-ore, ontology, owl, swrl, data-driven approach, provenance

Australian and New Zealand Standard Research Classifications (ANZSRC)

080608: Information Systems Development Methodologies

Contents

List of Figures

Chapter 1 Introduction

1.0 Background

In recent years, the amount of complex materials science data available from public databases [1-6], publishers [7-9] and within laboratories has grown exponentially. The databases and publishers provide high quality, scientifically-valuable data associated with both elemental and compound materials, while the laboratories hold experimental data generated by high-throughput laboratory instruments and data-intensive computational tools. First, this chapter discusses the emerging opportunity and challenges posed to materials scientists, because of the availability of such a wealth of widely-distributed materials science data and information. Second, this chapter also presents a case study, on which the requirements and the evaluation of a proposed technological solution are based. Third, the technical challenges of implementing the user requirements are identified and presented. Fourth, the thesis and objectives of this project are detailed. Fifth, the brief evaluation of the solution prototypes is described. Sixth, the previous related work of the technologies used in this project is examined. Finally, the thesis structure is outlined.

1.1 New Opportunity

The emerging wealth of complex materials science data and information has provided materials scientists with an alternative research approach—the data-driven science [10]. The basic distinguishing feature of the data-driven approach involves the attempt to detect actual patterns in the data and to use this information to make inferences to further patterns and/or to the causal mechanism underlying these phenomenological patterns [11]. This approach would be quite promising in reducing the number of costly experimental programs and shortening the lifecycle of the discovery of novel compound materials, in contrast to the traditional trial-and-error approach [12-15]. It relies on the old tactics of serendipity, craft knowledge and rules-of-thumb to modify existing materials, thereby making the discovery process very slow, costly and arduous.

Inspired by the exemplary projects in Bioinformatics [16, 17], Astroinformatics [18], Geoinformatics [19], and Biodiversity Informatics [20], materials scientists have started the data-driven approach that exploits a deluge of materials science data and information to optimise the materials design. Currently, they search across disparate, but closely related, public data sources, retrieve and integrate heterogeneous materials data, correlate acquired data with in-house experimental data, identify the predictive trend and common pattern, deduce a set of potential experimental parameters, validate them with computing modelling and, hopefully, pinpoint a set of potential parameters for setting up experimental programs. As a result, this approach could reduce the duplication of costly compound

preparation, testing and characterization. However, this approach also poses unprecedented challenges to materials scientists in materials science data management. The materials science community has called for a new discipline —Materials Informatics [15] for resolving these challenges.

1.1.1 Materials Informatics

Materials informatics is emerging as a new discipline to resolve the issues of data management, curation, integration and analysis that are challenging the materials scientists. Materials informatics is defined as the high-speed robust acquisition, management, analysis and dissemination of diverse materials data. Materials data access, acquisition, interoperability and curation were recently identified as being critical cyber-infrastructure imperatives for the materials science community [21, 22].

Critical requirements include the following. First, persistent unique identifiers for materials science resources are required. The resources are scattered around on the Web, but their discoverability is unreliable through the associated URLs [23], because the resources may be moved, removed and renamed for a variety of reasons. Thus, unique, standardised and persistent identifiers are required to continue to provide access to the resources, such as the Digital Object Identifier (DOI)[24]. Second, metadata standards are required for describing samples, processes and properties. The metadata standards enable materials scientists to easily understand, share and query materials science data. Additionally, they also facilitate data exchange between software applications, data transmission on the Web, and data processing. Third, the Web Ontology Language (OWL) ontologies are required for the correlation and integration of diverse, but related, materials science data. They provide:

- rich machine-readable semantic description –well-defined semantics of defined terms for computer programs to process information represented [25]
- formal definitions of domains by defining classes, properties and relationships between them
- a basis to enable, (i) matching between database schemas through the mapping between the elements of two schemas that are equivalent semantically and,(ii) reasoning and deduction of new information via the Semantic Web Rule Language (SWRL) rules [26] and a rule-inference engine, such as Pellet [27]
- Therefore, OWL ontologies enable semantic interoperability between resources, services, databases and devices via inter-related knowledge structures.

Fourth, a scientific workflow management system is required. Scientific workflows have emerged and been adapted from the business world as a means to formalize and structure the data analysis and

computation on the distributed resources. Because the process of compound preparation, testing and characterization is well-defined and repeated, ideally, it can be converted into a scientific workflow that coordinates, orchestrates and streamlines the experimental activities operated by both the multidisciplinary scientists distributed over a wide geographical region and the self-contained computational services within the compound synthesis process. Additionally, huge amounts of diverse and complex data products are generated from each phase of the scientific workflow. It is self-evident that paper-based laboratory notebooks are no longer capable of managing such magnitude and the complexity of those data. As a result, a workflow system is needed to manage the used and generated data, metadata and provenance information systematically. Finally, a system for publishing scientifically-valuable datasets is required. Scientific data dissemination, the last but important step in the lifecycle of the data-driven approach, enables the scientific results to be reviewed, verified and validated and reproduced, and for teaching and learning purposes.

1.2 New Challenges

In this section, the challenges confronting the materials scientists are identified and discussed. This project mainly investigates the challenges in data assimilation and dissemination. On one hand, data assimilation is about integrating diverse but related materials science data from disparate databases and about the management of huge amounts of complex data produced by the scientific experimental workflow. On the other hand, data dissemination is about the access, authoring and publishing of scientific data in standardised formats that are easily consulted and analysed, openly accessible and interoperable.

1.2.1 Materials Data Assimilation

Materials scientists search, retrieve and integrate data across publicly-available online databases and publishers to correlate the data with in-house experimental data for identifying trends, clusters and anomalies among the disparate, but closely related, data. When searching across autonomous data sources, materials scientists have to deal with different user interfaces and resolve inconsistent metadata terms, data structure, formats and metrics manually. For example, *temperature factors* can be represented in three different formats (Isotropic Temperature Factor (ITF), temperature factor (β) and the mean square amplitude of vibration (U)). As a result, this assimilating process is time-consuming and arduous, and could result in incorrect data collections. Following the data acquisition process, materials scientists start to investigate the correlations between processed data collections and the in-house experimental data.

In-house experimental data includes huge amounts of multivariate, multidimensional and mix-media data generated by the sophisticated and high-throughput laboratory instruments and data-intensive

computational tools from the experimental, characterization, testing and post-processing steps associated with the search for novel compound materials. Materials science data ranges from complex compound preparation and processing workflows to spectrographic analyses, 2D nano-scale microscopy images, textual publications, numerical data, animations, 3D crystallographic structures and complex phase diagrams. Those instruments include combinatorial and robotic laboratory instruments and atomic resolution microscopes, while the computational tools include high-performance modelling and simulation software tools. However, the wealth of diverse materials data is scattered around disparate physical and/or electronic storage systems, including laboratory notebooks, desktops, laptops, instruments and institute repositories, as a result of an impromptu, manual, unsystematic approach to data acquisition. Consequently, this approach not only impedes data access and sharing, but also could incur data loss, thereby resulting in a costly experimental duplication.

Additionally, the uncertainties of data quality are another issue to deter materials scientists from reusing in-house experimental data that has no associated provenance [28]. Provenance is essential within science, because it provides a history or documentation of the steps taken during the scientific discovery process. Understanding the source of data or how scientific results were arrived is essential for verifying or trusting that data and to enable its reuse and comparison. Precise, authenticated provenance data reduces duplication and insures against data loss, because the additional contextual and provenance information ensures the repeatability and verifiability of the results.

In addition to the challenges identified in the data assimilation, scientists across many domains, including materials science, are under increasing pressure from funding agencies and prestigious online publishers to publish experimental and evidential data together with the related traditional scholarly publications [29-32].

1.2.2 Materials Data Dissimilation

Data dissimilation is the last, but important phase in the data-driven approach for data sharing, reuse and e-learning. Currently, traditional scientific publications remain the predominant method for scholarly communication. In general, they present findings with the relevant methodology and evidential data for review and sharing. However, it is self-evident that traditional publications are not able to incorporate experimental data with the provenance information generated or used at every phase of the scientific investigation process to enable peer scientists to review, justify and repeat the findings effectively. As a result, the funding organizations, particularly in developed countries, are encouraging or even mandating the publication of scientific data together with the traditional

publications across many domains [29-32]. But there are a number of barriers that need to be overcome to encourage scientists to publish their datasets. These include, (1) poor accessibility to experimental data resulted from an impromptu, manual, unsystematic approaches to data acquisition and the lack of interoperability among the physical and electronic storage systems, (2) a lack of simple tools for publishing data with the provenance information, (3) a lack of motivation for scientists to spend time and effort preparing their data for publication, (4) concern with intellectual property rights and, (5) a lack of standards for publishing datasets with the provenance.

A number of different approaches have been implemented that link raw scientific data to scientific publications. Some online publishers, including Acta Crystallographica Section E – Structure Reports Online [33], Nature [7] and the ePIC Earth System Science Data and Methods [34] journal, enable the association of supplementary datasets with scholarly papers. Murray-Rust and Rzepa proposed the concept of datuments [35] — XML documents, that combine the data and the document using formal markup to allow processing and rendering in different ways via the Extensible Stylesheet Language Transformations (XSLT) [36]. A number of initiatives and projects (for instance, the Protein Data Bank (PDB) [37], CombeChem [38], eBank [39], NCBI [40], Virtual Observatory [41] and GBIF [42]) have spearheaded the development of infrastructures that facilitate the online publication of electronic scientific datasets. Although these existing approaches have advanced the publication of scientific data, they also have a number of limitations, including:

- The relationships between the datasets and publications are one-to-one, relatively fixed and involve hyperlinks with little or no support for semantics or provenance information
- A lack of flexibility or extensibility — scientists require simple, interactive GUIs that enable them to interactively define a set of resources generated from an experiment or investigation, relate them to each other and publish the lot as a package.
- The difficulty of discovering and retrieving components that are deeply embedded or hidden within HTML pages or the deep Web [43].

1.3 Case Study

Increasing demand for power and growing concerns about global warming has accelerated the search for sources of clean energy. Solid Oxide Fuel Cells (SOFCs) [44] have been widely identified as highly-efficient electrochemical energy converters with a reduced production of greenhouse gases. However, they are reliant upon new or improved materials being discovered to make them economically viable and sufficiently reliable, durable and efficient for both mobile and stationary power generation.

The core component of an SOFC is an oxygen-ion-conducting membrane that must have mechanical and chemical stability at elevated temperatures under both reducing and oxidising conditions. Although oxygen-ion-conducting materials with these properties have been known for approximately 100 years [45], there is a demand for new, structurally stable materials with **oxygen conductivities approaching > 10^{-1} Scm^{-2}(Siemens per centimetre squared) at temperatures below 500°C**. The present solid oxide fuel cell systems operate at temperatures in excess of 750°C; a reduction in operating temperature has a significant effect on the cost, durability and engineering demands of the ancillary materials that make up the device. Being able to operate at 500°C brings in the possibility of metal manifolding and much simpler heat management systems. Figure 1.1 [46] illustrates the operating range of the current oxygen ion conductors. The shaded triangle indicates the area where we are intending to discover new compounds.

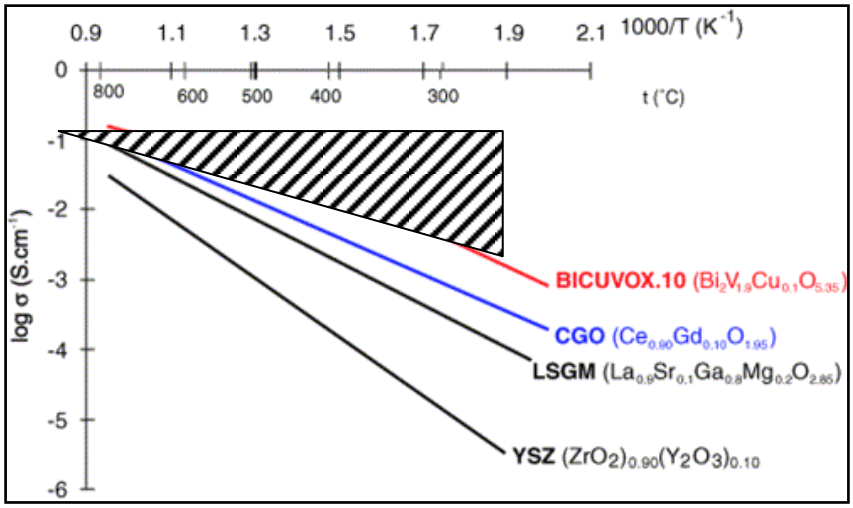


Figure 1.1: Ionic Conductivity of Current Oxide Ion Conductors as a Function of Inverse temperature [46]

Fuel-cell scientists from the Australian Institute for Bioengineering and Nanotechnology (AIBN) within the University of Queensland believe that the data-driven approach [15] would expedite the design and discovery of the novel oxygen ion conducting materials. The approach is to combine novel data management [47], data mining [48] and computational modelling techniques [49, 50] to interrogate the wide range of experimental data available — to systematically discover new materials with desirable properties. The AIBN scientists' arguments are:

- A substantial body of high quality experimental data associated with both elemental and compound materials has been captured over the past 10–20 years. This is now available in publicly accessible, comprehensive databases that contain crystallographic structure data, thermodynamic data, phase stability data and ionic conduction data.

- Databases, data retrieval, integration and mining techniques, high performance computing, algorithms and computational modelling have improved significantly in recent years. These advances will enable the correlation of large data sets from different experimental and characterization techniques, improved accuracy of simulations and predictive models and, automatically-generated experimental workflows.
- The stringent criteria required by solid oxide ion conductors for fuel cell applications significantly narrows the search space for possible compounds. This greatly improves the potential for positive results from predictive data mining and computational modelling.
- The ABIN scientists have already, intuitively identified underlying patterns that indicate new fertile areas for further searching [51-53]. But the size of the datasets prohibits the traditional trial-and-error experimentation involving costly compound preparation, characterisation and testing. Virtual screening and ab initio validation can significantly reduce the amount of experimentation and the associated effort and costs.

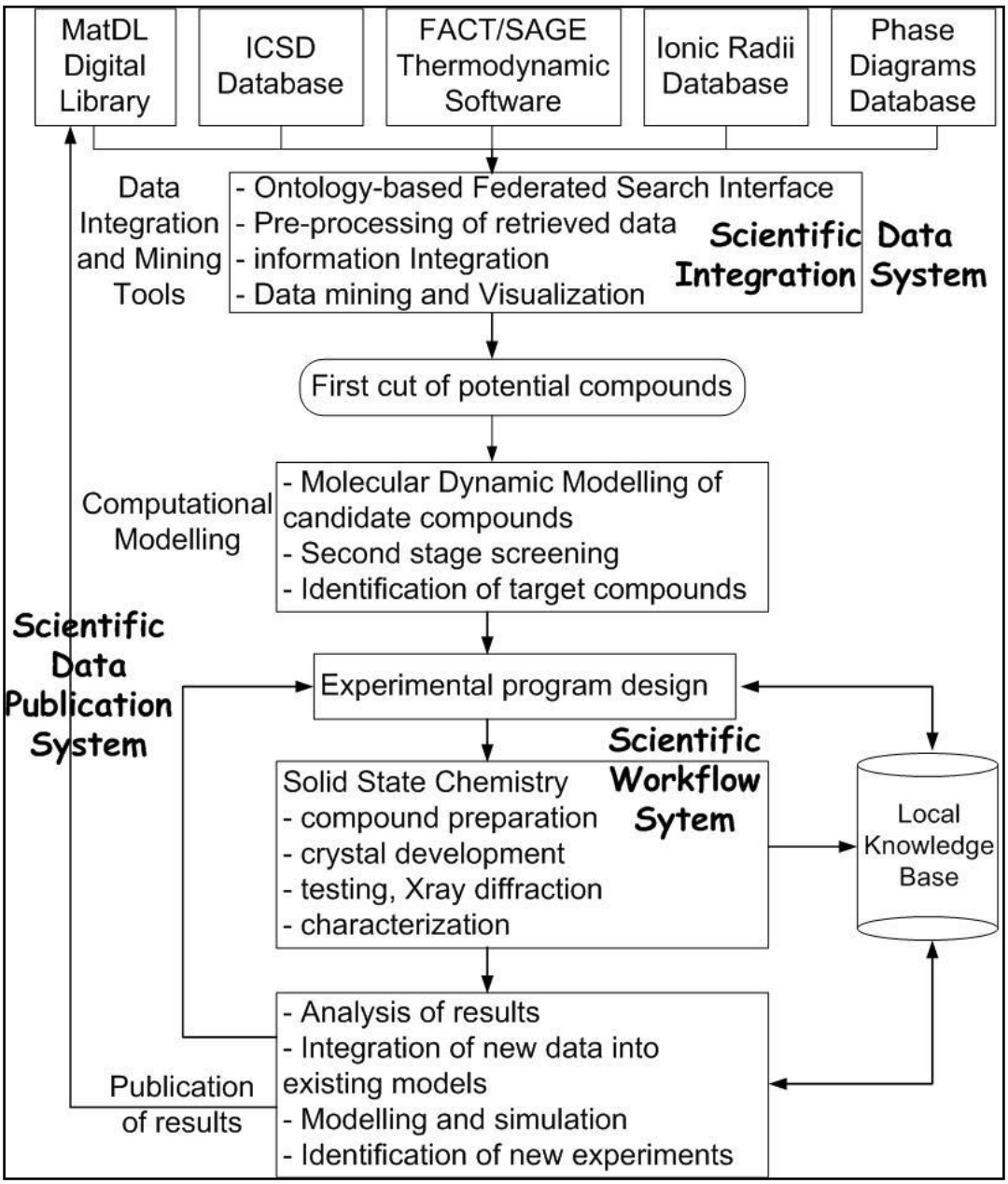


Figure 1.2: Schematic Illustration of the Overall Project Structure and Components

As a result, the AIBN scientists demanded a full-fledged Materials Informatics Workbench that enables them to, (1) search across disparate databases and accurately correlate and integrate diverse but related materials science data for further analysis and modelling, (2) effectively coordinate a complex compound synthesis program involving the collaborating scientists across multiple organisations and computing services, and systematically and precisely capture and store the used and generated data, metadata and provenance into distributed databases and, (3) easily access, author and publish scientific datasets for widespread dissemination.

Figure 1.2 demonstrates the schematic overview of the overall initiative. This project focuses on the development of the Materials Informatics Workbench in the aspects of scientific data integration, scientific workflow management and scientific data publication.

1.4 Technical Challenges

The complexity of the AIBN scientists' requirements poses non-trivial challenges to the development of the proposed Materials Informatics Workbench. The following details these challenges associated with the implementation of the workbench components shown in Figure 1.2.

The AIBN scientists required a user-friendly, easy-to-learn user interface that enables them to search with keywords across disparate databases and retrieve, correlate and integrate diverse but related data for further analysis. There are two identified major challenges to be overcome. First, through the search keywords, the potential system should be able to identify which databases are involved in the search request. It should also be able to construct the query statements dynamically and accurately for querying the databases. Second, the system should be able to resolve the semantic differences of the metadata terms associated with the data items retrieved from those databases to correlate and integrate the retrieved data items correctly.

The scientists also require a system that enables them to, (1) automate and streamline the compound synthesis process involving cross-disciplinary scientists across multiple organisations and self-contained computational resources in a pre-defined order and, (2) acquire data produced at each stage of the process. The potential system should be underpinned by a proper, reliable workflow engine. The engine should be able to run a workflow instance that coordinates, orchestrates and interoperates the human and computational activities seamlessly. The allocated participants will be alerted, once the assigned tasks are available. The computational resources include the database systems for data storage, such as MYSQL, Application Program Interfaces (APIs) such as the JavaMail API for alerting the participants by emails and analysis tools such as MATLAB for image analysis. For an example of the image analysis through MATLAB, the runtime workflow instance should invoke MATLAB and pass an image to MATLAB as input. Then, without human intervention, the workflow instance should enable the capture of the analytical result generated by MATLAB and the storing of the result to a designated storage system.

Finally, the scientists require a tool that enables them to easily access, author and publish their data with contextual and attribution information, with measures for the IP protection and in the formats enabling widespread dissemination. These requirements are so complex and technologically demanding that no single technology is able to provide a full and effective solution for them. The

resulting challenges confronting this project are to evaluate, extend and combine proper technologies for satisfying these requirements.

Rendering a view of a visualized scientific experiment workflow provides an intuitive, ideal way for materials scientists to access data. The visualized scientific process consists of nodes and arrows. The former represents either experimental data or activities, while the latter indicates the relationships between the nodes. This view enables scientists to discover and access the information for the experimental components, such as experimental data, activities and equipment settings, because scientists are familiar with the structure of the scientific investigation process. However, scientific workflows are complex and the sub-workflows are sometimes conducted by cross-disciplinary scientists. As a result, it would be quite challenging for materials scientists to understand and comprehend the entire workflow. Additionally, the collaborating scientists tend to keep their experimental information and data from others. The technical challenge confronting this project is how to ensure viewers can easily understand the science of highly complex workflows and guard the confidential information properly.

Materials scientists want to author a data package that encapsulates the scientific methodology with the essential datasets and contextual information to support their scientific contribution comprehensively. However, scientific discovery workflows are complex. The technical challenge is how to enable scientists to easily build a concise, coarse-grained view of a scientific methodology, including the essential datasets and contextual information, from the comprehensive, fine-grained view of the scientific workflow provenance trail. Additionally, they also want to incorporate published scholarly papers and/or peer-reviewed datasets to support their findings.

While a data package is being wrapped up, materials scientists wanted to, (1) enrich the package with contextual information to be self-contained and explanatory, (2) enforce measures for the intellectual property protection, (3) ensure the package is in the formats that maximize its dissemination — easily discoverable on the Web by both machine and human agents and, (4) incorporate the package into a digital library. These requirements give rise to a number of technical challenges, (1) investigating and selecting the proper metadata schemas to specify the contextual and attribution information, (2) evaluating and selecting the proper Web-based copyright license to protect the data package's IP, (3) investigating and selecting a proper publishing standard that enables the data package to be machine- and human-readable, interoperable, exchangeable and reusable, (4) converting the data package to the Web publishing formats supported by the potential publishing standard and, (5) investigating and selecting a digital library with a high popularity and developing a plug-in to convert the data package to the library's ingesting format and export it into the library directly.

This project aims to develop prototypes for the major components of the Materials Informatics Workbench by applying, evaluating and extending the emerging, novel, innovative open-source technologies. Hopefully, these technologies will resolve the technical challenges identified above and provide the materials scientists with an effective informatics solution to expedite the discovery of the novel oxygen ion conducting materials.

1.5 Overall Objectives

This thesis aims to apply, evaluate and extend the emerging technologies and specifications of the Semantic Web, the Web Service Business Process Execution Language (WSBPEL) [54] and Open Archives Initiative Protocol – Object Reuse and Exchange (OAI-ORE) [55] to provide innovative approaches, in the context of materials science, to, (1) semantically correlate and integrate distributed, but related, diverse materials science data and information, (2) manage the scientific investigation process and data products through the WSBPEL workflow and, (3) publish scientific results as a scientific compound object [56] complying with OAI-ORE that makes the information within the object machine-readable, discoverable, sharable and reusable.

This approach has ingeniously combined and extended a number of state-of-the-art technologies that enable the development of novel user interfaces and innovative algorithms and techniques behind the major components of the proposed Materials Informatics Workbench. As a result, it supports the faster, more accurate assimilation and dissemination of materials science data. In particular, this thesis focuses on providing:

1. The Materials Science Ontology that enables
    - the mapping between and the integration of disparate materials science databases
    - the modelling of experimental provenance information captured in the physical and digital domain
    - the inferencing and extraction of new knowledge within the materials science domain.
2. An ontology-based federated search interface that enables materials scientists to search, retrieve, correlate and integrate diverse, but related, materials science data and information across disparate databases
3. A scientific workflow management system that manages:
    - the scientific investigation process that incorporates cross-organization scientists and self-contained computational services
    - experimental data and information generated by the process.

4. A provenance-aware scientific compound object publishing system that:

   - provides scientists with the view of the highly complex scientific workflow at multiple-grained levels, so they can easily comprehend the science of the workflow, access experimental data and information and keep the confidential information from unauthorised viewers
   - enables them to quickly and easily author and publish a scientific compound object that:
     - incorporates the internal experimental data with the provenance information from a visualised scientific workflow and external digital objects with the metadata discoverable via the Web
     - is self-contained and explanatory with Intellectual Property protection
     - is ensured to be disseminated widely on the Web.

Overall, this will enhance the data-driven approach to expediting the discovery of novel compound materials.

1.6 Evaluation

A test case from the materials science research described in Section 1.3 has been used to evaluate the effectiveness and usefulness of the prototypes for the major workbench components. The three sets of questions below have been formulated accordingly for user feedback for each workbench component, respectively. The prototypes that have been developed have been assessed by the AIBN scientists against the following questions that the user feedback follows:

1. *How well does the federated search interface underpinned by the Materials Science ontology support materials scientists to search, retrieve, correlate and integrate diverse but related data across materials science databases? What are the limitations?*

The AIBN scientists were impressed with the convenience of the system because of the easy and quick access to the integrated databases and analytical tools. However, one of the major limitations identified was the lack of data in the publicly-available databases that we have incorporated. Commercial databases are more complete and comprehensive, but outside the scope and budget of this project. Hopefully over time, the culture of sharing materials science data through open-access archives will become more widely adopted in the materials science community and this situation will improve. The other limitation is that adding new databases requires human effort to match the databases schemas through the underlying ontology and populate the names of databases, entities

and attributes as instances into the ontology from the schemas. Ideally, the uploading and mapping of new database schemas could be streamlined via a Web interface.

2. *How well does the WSBPEL workflow management system support the collaboration between the collaborating scientists and computational services within the scientific discovery process? Has the system met materials scientists' requirements in data capture? What are the limitations?*

Following the trials, the AIBN scientists experienced a surge of relief that the complex compound synthesis process could be managed by the WSBPEL workflow system. The system can also acquire the huge amounts of complex data generated by the computational service that is no longer handled manually with paper-based laboratory notebooks. The scientists can monitor the progress of the running workflow via a Web interface. Amazingly, the scientists discovered that a scientific workflow encapsulates scientific intellectual property and enables the sharing of knowledge between scientists. However, one of the major limitations was identified in that they are required to input the parameters of equipment settings for the laboratory instruments manually. Ideally, the equipment settings should be fed into laboratory instruments directly, without human intervention. The other limitation was that the workflow script was pre-defined and fixed. Materials scientists would like a scientist-friendly graphical workflow editor in place to design their own workflows.

3. *How well does the proposed publishing system overcome the barriers identified in Section 1.2.2 for scientific data publication? How well are the user interfaces for materials scientists to author and publish scientific datasets? What are the limitations?*

Following the usability testing, the AIBN scientists were very positive. First, they were presented with a high-level view of a highly complex workflow and then they were able to expand the view gradually for more details according to their access privileges. As a result of the multiple-grained level-of-view approach, the users can zoom in and out the workflow at will, while the confidential information can be kept away from unauthorized users. Second, they were able to graphically link internally-generated provenance trails to external resources, discoverable via a Web browser. This allowed authors to include other relevant research outcomes to strengthen the claims of their findings, thereby making their research outcomes more comprehensive but still self-contained and facilitating the peer-review process. Third, they were allowed to interactively generate coarse-grained views of scientific experimental processes via automatic inferencing. Fourth, they were able to attach attributions in the Dublin Core format and the Creative Commons license for attribution and IP protection, respectively. Finally, they were able to publish scientific datasets as a scientific compound object, complying with OAI-ORE in a variety of Web publishing formats that maximize

the dissemination of the datasets over the Web. However, the system does not support searching, reloading, and editing of published OAI-ORE scientific compound objects.

The following examines the previous, related work for the major components of the proposed Materials Informatics Workbench.

1.7 Related Work

This chapter examines previous, related work in the fields of materials science data integration, scientific workflow management, scientific provenance visualization and scientific data publication.

1.7.1 Materials Data Integration

The barriers to data acquisition within materials science [21] have been identified. They include locating the correct data sources, retrieving required data with the metadata, correlating and integrating information retrieved from different data sources and storing the information into a fully documented, searchable and accessible storage system. Materials scientists demand: a fully-fledged tool help them streamline the process of selecting data sources adapting different user interfaces; retrieving required data with the metadata; resolving differences in data formats, structures and metrics; mapping different metadata terms; and correlating and integrating retrieved information accurately. The following describes the previous data integration approaches, including data warehouses and mediated schemas, adopted within the materials science domain.

1.7.1.1 Data Warehouse

The data warehouse approach [57], proposed by Li, aims to resolve the issues of the low precision and recall of the retrieved data during the materials selection processes. The approach extracts data from databases, identifies and resolves data differences and loads the cleansed data into a repository. As a result, it provides a single platform for the end-users to search data more quickly and easily, improves the recall and precision rates and ensures the availability of deposited data. However, this approach is inflexible and cannot easily adapt to changes in database schemas. Additionally, it incurs costly overheads for data updates, cleansing and maintenance. These limitations led to the development of a more flexible approach, based on mediated schemas.

1.7.1.2 Mediated Schema

In contrast to the above tightly-coupled approach, the mediated schema is loosely-coupled and cost-effective. It is implemented in two ways, global-as-view and local-as-view [58]. The former requires that the global schema is expressed in terms of the local schemas, while the latter requires that the local schemas are defined in terms of the global schema. In this review, I focus on the local-as-view approach with either the XML schema or the OWL ontology as the global schema, because this

approach not only reconciles the differences in the data structures between the local sources, but also makes no changes to the global schema while adding or removing liaised local sources. Furthermore, the ontology approach enables the mapping of the metadata terms between the global and local schemas; thus, the metadata terms are not necessarily identical between both schemas when adopting the local-as-view approach.

1.7.1.2.1 XML Schema

There have been a number of initiatives on the development of XML schemas within the materials science and relevant domains, such as chemistry, including the Chemical Markup Language CML, Materials Property Data Markup Language MatML and the Gas Hydrate Markup Language GHML.

CML [59-65], the first domain-specific implementation based strictly on XML, is developed for containing chemical information components within documents. Its design supports interoperability with the XML family of tools and protocols. It is capable of supporting a wide range of chemical concepts including molecules, reactions, spectra and analysis data, computational chemistry, chemical crystallography and materials.

MatML [66] is an extensible markup language developed especially to facilitate the materials data interchange between applications and data transmission over the Web in response to search and efficient data processing. It can uniformly represent materials property data to resolve syntactic and structural heterogeneity. Because MatML is simple, flexible and understandable, it offers many benefits to materials scientists and engineers [67].

The CODATA Hydrate Task Group has developed an XML schema for the Gas Hydrate Markup Language GHML [68, 69] for resolving differences in metadata terms and data structures between gas hydrate databases. XML Schemas provide a shared vocabulary and formalize constraints in data structures and types.

However, XML schemas embody a tree structure (acyclic graph) that has monotonic, implicit and ambiguous relationships between connected nodes. As a result, XML Schemas provide little support for the semantic knowledge necessary to enable flexible dynamic mapping vocabularies. In contrast to XML schemas, OWL ontologies enable semantic mapping between domain-specific knowledge structures by declaring binary relationships between the nodes, and inferring potential relationships between the nodes via reasoning engines. Within the materials science community, there have been a number of previous efforts at the development of ontologies to resolve the different issues, including data integration.

1.7.1.2.2 Semantic Web

Ontologies are an integral component of the Semantic Web. A variety of ontologies have been developed to, (1) extract knowledge from text, such as PLINIUS's ontology [70], (2) model a cross-disciplinary classification scheme for nanoscale research, such as Tanaka's meta-level ontology [71] and, (3) solve the data integration issues, such as Ashino's Material Ontology [72, 73] and Zhang's Semantic Model for Materials Scientific Data SMM [74]. The following clarifies the effort at applying OWL ontologies to materials data integration issues.

Ashino developed the Material Ontology [72] that consists of seven modules, Materials Information, Substance, Property, Environment, Process, Unit Dimension and, Physical Constant. The vocabularies are from two prestigious sources, Matdata.net [75] and NIMS's MatNavi [76]. Such an approach could lower the ontological commitments [77]. *Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner*. Because agents complying with those vocabulary sources can share the Material Ontology, they can selectively commit to the individual modules rather than the entire ontology. Thus, these two factors greatly contribute to the lowering of the ontological commitment. However, the structure of the Property Module is questionable. For example, the property ontology file indicates that *property:PhysicalQuantity* is subsumed to *property:Property*. The subclass's name suggests it is a quantitative indicator and is not describing material characteristics like its sibling classes, such as *property:Mechanical and property:Electrical*.

Zhang's group [74] has developed two types of ontologies, the domain-specific and the mapping ontologies. The domain-specific ontologies include the global and local ontologies. The global ontology encapsulates the high level structure of the materials science knowledge, named the Semantic Model for Materials scientific Data (SMM), while the local ontologies are based on local database schemas. The mapping ontologies define the mappings between the SMM ontology and the local ontologies, and between the local ontologies and the database schemas. However, the authors do not discuss how they ensure the SMM quality, particularly in coherence [77]. They developed inference rules for defining new concepts, such as *CorrosionResistantMaterial*. In addition, they developed a mapping ontology — a data source description ontology OWL-DSDO for structuring the names of databases, entities and attributes from the domain-specific database schemas at the instance level. They then mapped the local domain-specific ontologies to the OWL-DSDO ontology at the class level through the declared object properties *RelatedOntClass* and *RelatedProperty.* However, those ontologies are in different worlds, the local ones are about materials science, the others about relational database schemas. Thus, such class and property mappings are semantically mismatched.

My approach is similar to Zhange's SMM ontology and Ashino's Material Ontology. I have developed the Materials Science ontology described in Chapter 3 for encapsulating the high level knowledge of materials science. However, I adopt a different mapping approach from Zhang's, which is described in Chapter 4.

1.7.2 Scientific Workflows Management System

Scientific workflows technologies [78] represent an increasingly important component of the scientific discovery process. However, the adoption of scientific workflows in the materials science community remains in the infant stage. I have investigated a number of the commercial and research workflow management systems used by the materials science community.

SciTegic's Pipeline Pilot [79] and InforSense's Knowledge Discovery Environment KDE [80] are commercial products that are data pipelining workflow systems. Data pipelines [81] are a specific form of workflows and focus on the management of data records. Data pipelines enable processing data to move between tasks rapidly, thereby minimising memory footprint and disk access and optimizing data-throughput rates. Both systems streamline and automate the process of data retrieval, organization, analysis and reporting. Additionally, they provide visual editors for wiring workflow components together graphically. SciTegic's Pipeline Pilot with the materials component collection [79] automates the processing of powder diffraction data and characterizes compounds in terms of their molecular structure, chemistry, and structural properties.

InforSense's KDE has been incorporated as an integral part of an informatics platform [82] developed for TOPCOMBI [83]. The authors [82] demonstrate the ability of workflow paradigms in achieving time-effectiveness and data traceability through two case studies, high-throughput kinetic on glycerol oxidation and the virtual screening of catalysts. InforSense's KDE is a commercialised product of DiscoveryNet [84] that is a suite of software tools built on top of the Uniform Interface to Computing Resources (UNICORE) [85]. Workflows composed through InforSense's KDE are represented in Discovery Process Markup Language (DPML). DPML supports both a dataflow model of computation for analytical workflows and control flow operations for linking and orchestrating multiple analytical workflows. UNICORE provides a seamless interface for preparing and submitting jobs to a wide variety of heterogeneous distributed computing resources and data storages. It supports users running scientific and engineering applications in a heterogeneous Grid environment. In addition to the commercial products, there are also research projects solving similar issues, including the CDK-Taverna and ASKALON projects.

The CDK-Taverna project [86] is an open-source workflow solution through the combination of Taverna [87], the Chemistry Development Kit (CDK) [88, 89] and Bioclipse [90]. First, Taverna is a

workflow workbench that allows bio-informaticians to construct high-level workflows that integrate Web services including molecular biology tools and databases, into a single analysis. The workflows are represented in Simple Conceptual Unified Flow Language (Scufl) [91] that enables Taverna to link and orchestrate web services, and to model a network of processing activities linked by data and control flows. Next, the CDK component provides a Java library of more than 100 components for Structural Chemo- and Bio-informatics including 2D and 3D rendering of chemical structures, the Simplified Molecular Input Line Entry Specification (SMILES) parsing and generation, and the Quantitative Structure-Activity Relationship (QSAR) descriptor calculations. Additionally, the CDK component has been integrated with R — an open-source free software environment for statistical computing and graphic — for enhancing the utility of the statistical environment for chemo-informatics applications. Finally, Bioclipse is an Eclipse-based viewer for demonstrating the results of data analysis in different formats simultaneously. The results are about either molecular analysis from the chemo-informatics perspective or sequences/protein analysis from the bio-informatics perspective. The CDK-Taverna project aims at creating a flexible, extensible and easy-to-use system that allows the construction of powerful data workflows in a Lego-like manner. Potential workflows resolve data filtering, migration and transformation, information retrieval, Quantitative Structure-Activity Relationship (QSAR)/Quantitative Structure-Property Relationship (QSPR) or pharmacophore-related tasks, data analysis (statistics, clustering and computational intelligence), analytical and spectroscopical support, and molecular modelling.

The ASKALON [92] project intends to make the Grid environment transparent to end-users, and provides them with a UML-based visual editor, through which target users can compose and model workflows. They can then send them to the ASKALON Web Services Resource Framework-based middleware services for scheduling and reliable execution on Grid infrastructure. The workflows are represented in the Abstract Grid Workflow Language (AGWL) [93]. AGWL can represent data flows and control flows including sequence, parallel, iteration and choice. The authors illustrated the ASKALON project with a materials science case study. They ported WIEN2k [94] as a Grid application by splitting the monolithic code into several coarse-grained activities coordinated in a workflow. However, a number of drawbacks of the above-discussed systems have been identified.

Even though the reviewed systems enable scientists to collaborate computing resources for data analysis more easily, there remain issues to be rectified. First, the workflow languages are not standardized, so the resulted workflow scripts are incompatible with other workflow systems. Second, the systems lack formal mechanisms for supporting the collaboration, including human participation during workflow execution. Third, except for the CDK-Taverna, they do not have mechanisms for supporting the capture and access of experimental data and provenance information

generated during the workflow execution. Fourth, even though ASKALON and CDK-Taverna have fault tolerance mechanisms, they do not support backwards error recovery — the ability for undoing specified completed work. Finally, their sustainability is uncertain. The further development of the research projects depends on the availability of highly sought-after research funding, while the Research and Development of the commercial products relies on their firms' profitability.

The main motivations for selecting the Web Service Business Process Execution Language (WSBPEL) [54] underlying the workflow management system discussed in Section 2.5 are the following. First, Akram et al. [95] argue that WSBPEL would be the best incumbent candidate for scientific workflows, if complemented with the Web Services Invocation Framework [96], J2EE and WS-* specifications. This results from the evaluation of the WSBPEL features against the distinguishing characteristics of scientific workflows, including modular design, exception handling, compensation handler, adaptability and flexibility and the management of the workflow. Next, the WSBPEL workflows are more sharable, reusable and sustainable in contrast to the reviewed workflow systems, because WSBPEL is proposed and endorsed by industry giants, such as Microsoft, IBM, SAP and so on. They are also backed by the open-source communities through the provision of commercial quality workflow engines, such as the ActiveBPEL [97] and the Apache Orchestration Director Engine (ODE) engines [98], and by visual design tools, such as Eclipse's BPEL designer. These tools are gradually being adopted by the research communities [95, 99, 100]. Then, the WSBPEL system is easily extended to work within the Grid environment through the implementation of the Open Grid Services Infrastructure and WS-Resource Framework [101, 102]; the extended WSBPEL system is provenance-aware [99, 103] and provides a formal mechanism for incorporating human interaction in the workflow environment through BPEL4People [104]. Finally, WSBPEL has a well-defined specification for handling fault-tolerance and compensation.

1.7.3 Provenance Visualisation

Visualizing scientific provenance trails provides scientists with an intuitive view of the scientific discovery process, so the scientists can easily review, verify and validate, and reproduce scientific results. Even though there have been different approaches to visualising provenance data stored in relational or RDF-based storage systems, the approaches do not support the rendering of multi-granularity levels of views for different purposes. The coarse-grained view serves for teaching or publication purposes and protects confidential information or intellectual property, while the fine-grained view details experimental information. The following describes a previous research effort at provenance visualisation.

The *Prototype Lineage Server* [105] allows users to browse lineage information by navigating through the sets of metadata that provide useful details about the data products and the transformations in a workflow invocation. Web server scripts on the lineage server query the lineage database and provide a Web browser interface that allows navigation via HTML links. Views are restricted to parent and children metadata objects. Clicking on a parent object will move that link to the centre of the screen and show that object's parents. Clicking on the metadata object link in the centre of the screen will bring up the XML metadata for an object.

Pedigree Graph [106], one of tools in the Multi-Scale Chemistry (MCS) portal from the Collaboratory for Multi-Scale Chemical Science (CMCS), is designed to enable users to view multi-scale data provenance. The portlet provides scientists with a two-dimensional visualization of a data object or file and all of its scientific pedigree relationships. The view is static and rendered straight from GXL (Graphical eXchange Language) files, but users are able to traverse the tree by clicking on links.

The myGrid project renders graph-based views of the RDF-coded provenances using Haystack [107]. This is used to visualize networks of semantic relationships among provenance resources associated with experiments. Haystack [108] is a Semantic Web browser that enables developers to provide tailored views over RDF-metadata. The authors point out that Haystack is highly resource-consumptive, because its execution is based on Adenine [109], a high level programming language developed on top of the Java Programming Language. Hence, the response time to the user's instructions could be slow.

The *VisTrails* system [110] was developed by the University of Utah for building, storing, editing and visualizing workflows and interactively tracking workflow execution and evolution. Although it uses graphs to visualize workflows and provenance trails, it is not designed to generate personalized views of provenance adapted for publication or teaching purposes or to suit a user's interest or access permissions.

There are existing systems that enable the visualization of RDF-encoded provenance graphs. However, the unique aspect of the Provenance Visualization component within the Scientific Compound Object Publishing and Editing (SCOPE) system, described in Chapter 5, is its ability to generate personalized views of the provenance relationships automatically, by using a combination of user input and access policies.

1.7.4 Scientific Data Publication

Publishing scientifically valuable datasets is the last but significant step in the lifecycle of the data-driven approach. However, current approaches have made scientists reluctant to publish their data

within the relevant traditional publications, even though they are under increasing pressures from funding agencies. This section describes the current approaches by which data is linked to scholarly publications:

1. The first method involves either including a reference from the paper to an accession number in a database or adding a hyperlink from the paper to a dataset or data held within a database via a unique identifier (for example, many publishers use Digital Object Identifiers (DOIs)).
2. The second approach involves embedding the data within the scholarly publication via a formal markup language.

Examples of publishers who support the first approach include, Nature [7] and the American Chemical Society [8]— which require that papers about proteins, DNA sequences or molecular structures must associate them with accession numbers assigned by designated publicly-accessible databases, such as Genbank [111], the Protein Data Bank (PDB) [37] and SWISS-PROT [112]. This approach depends on the long-term availability and accessibility of the large-scale online databases of scientific data.

The Protein Data Bank (PDB) [37] is just one example of a public database that is built from user submissions. Other similar large-scale online aggregated databases have been established and are maintained by organizations, such as NASA [113], NIST [114], NCBI [40], STD-DOI [115], GBIF [42] and NOAA [116]— for research domains including global atmospheric and climatic research, computational chemistry, genomics, earth sciences and analytical physics. Typically, these organizations collect data in their own database schema and if others want to upload their data, it must first be converted to the organizational database schema and specified formats. The problems associated with this first approach include:

- The link from the paper to the data is usually unidirectional and does not include any semantics or provenance information. Discovering the data via web crawlers is not possible, because it is part of the *deep web* [43].
- The procedure for submitting papers and/or data to online publishers and publicly-accessible databases is database-specific and rigid. Understanding those procedures can frustrate or de-motivate scientists from publishing their data.

The second approach to publishing raw data linked to publications involves using some form of XML to markup a scientific publication structurally and semantically — to distinguish between and to interpret the publication text and different types of embedded or related data. Examples of this approach include the Murray-Rust's datuments [35]— XML documents that are machine-readable

and can be rendered in different ways using XSLT. The eCrystals Crystal Structure Report Archive [117], a subproject under CombeChem [118] and eBank [39], publishes first-hand but non-peer-reviewed crystallographic data online. All information about a single crystal is dynamically generated as a highly structured web page with detailed provenance information and links to related citations. Acta Crystallographica Section E – Structure Reports Online [9] also binds hyperlinks to the paper and supplementary material under the one title. The German Scientific Drilling Database (SDDB) project [119] is also investigating the use of XHTML to integrate geological sample information with derived data and published studies in which the data is interpreted.

The major drawback associated with the second approach is that many Web spiders/crawlers cannot determine the semantic relationships between raw data and HTML text bound within a single web page. Explicitly typed relationships, as defined within Named Graphs [120], are required to raise the relationships between the components to first class objects that can have their own provenance information.

However, some of the limitations of the XHTML approach to scientific publishing may be overcome through the adoption of emerging technologies, such as Microformats [121], RDFa [122] and GRDDL [123]. Microformats and RDFa enable semantic tags to be embedded within XHTML documents to tag the content or link to related documents or data (via the *rel* tag) — without affecting the display of the HTML text. The inclusion of these light-weight semantic tags enables machine understanding, interpretation and processing of the publication. GRDDL (Gleaning Resource Descriptions from Dialects of Languages) is a mechanism for deriving formal metadata in RDF by using XSLT to process XHTML documents, to extract the embedded semantics (for instance, RDFa). The future may well see scientific publication authoring systems that use RDFa to embed tags or annotations in (X)HTML files and use RDF-a aware browsers or GRDDL to extract this, convert it to RDF, store it in an RDF triple store and search it using the SPARQL Protocol and Query Language (SPARQL).

My approach, detailed in Chapter 5, enables end-users to, (1) author publishing datasets by drag-and-drop, (2) attach contextual information to the datasets that is critical for data sharing and reuse [124] by associating metadata with individual datasets, specifying or inferring via a rule-reasoning engine the direct relationship between the datasets, and creating and attaching metadata for the resulting data package, (3) attach the Creative Commons license [125] for how the data package may be used, (4) publish the package as a scientific compound object complying with the Open Archives Initiative Protocol – Object Exchange and Reuse (OAI-ORE) in standardized web formats including ATOM 1.0 [126] that can be indexed by major search engines and discoverable subsequently over the Web and, (5) export the package to a Fedora digital library [127].

1.8 The Structure of the Thesis

Chapter 2 describes the different phases of this project and outlines the overall architecture and components of the Materials Informatics Workbench.

Chapter 3 presents the Materials Science Ontology (MatOnto) and its development. This ontology, underlying the MatSeek system described in Chapter 4, is used for mapping to the schemas of the materials science databases. Additionally, MatOnto not only enables a semantic layer between the provenance (relational) database and the SCOPE system, but also contains the SWRL rules to infer a coarse-grained view from the fine-grained lineage view of the visualised scientific experiment workflow, both of which are described in Chapter 5.

Chapter 4 describes the MatSeek system — a federated search interface, based on the MatOnto ontology (Chapter 3), to key materials science databases and analytical tools.

Chapter 5 describes the Scientific Compound Object Publishing and Editing system (SCOPE) that is developed to enable scientists to intuitively access, author and publish scientific data as scientific compound object complying with OAI-ORE in standardised web publishing formats.

Chapter 6, the concluding chapter, summaries the work done for this thesis, describes its results and significance and limitations, and possible improvements.

The appendices contain the MatOnto ontology in the Manchester OWL Syntax (Chapter 3), the WSBPEL workflows In the Business Process Modelling Notation (Section 2.5) and the examples of the D2R Map (Chapter 5).

Chapter 2 Project and Workbench Overview

This chapter describes the project background and structure, and the architectural overview of the proposed Materials Informatics Workbench.

2.0 Project Background

The Materials Informatics initiative discussed in Section 1.1.1 has been acknowledged as an increasing important component of the cyber-infrastructure [22]. Meanwhile, associated barriers [21] relating to data assimilation and dissimilation were also identified by the materials science community. The lack of tools for resolving those barriers has deterred materials scientists from adopting this new initiative.

Locally, at the University of Queensland, fuel cell scientists in the Australian Institute for Bioengineering and Nanotechnology (AIBN) also endorsed this initiative and proposed an exemplary project that applied an innovative data-driven approach to expediting the design and discovery of the novel oxygen ion conductors described in Section 1.3. They believed that the availability of rich, high-quality experimental data associated with both elemental and compound materials might enable them to pinpoint a set of potential parameters for generating very precise and targeted compound synthesis programs —thereby reducing the duplication of costly compound preparation, testing and characterisation.

The narrower scope of the problem and the availability of advanced local expert knowledge offered significant potential for an application of the Semantic Web, Web Service Business Process Execution Language (WSBPEL) and Open Archives Initiative Protocol – Object Exchange and Reuse (OAI-ORE) technologies as the base for the development of the proposed Materials Informatics Workbench. Materials science provided terminology that was well-defined, consistent and rich in detail. In the past few years, there have been huge amounts of high-quality materials science data and information available from online sources. Searching and processing, correlation and integration, analysis and modelling of the materials data potentially significantly reduced the amount of experimentation and the associated effort and costs. From the fuel cell scientists' perspective, this new approach was extremely valuable. In this way, fuel cell scientists would sharply shorten the cycle of the discovery of novel compound materials.

In addition, there was the opportunity for a greater measure of control on the production and analysis of experimental data and the systematic generation of high-quality, consistent metadata. Therefore, that provided the easy access to recorded experimental data for easing the scientists' burdens for publishing scientific results. A simple way to select, encapsulate, attach copyright licenses, such as

Creative Commons, and format scientific results in a standardized way would encourage scientists to publish their data. As a result, they spent less time and effort preparing their data for publication, had less concern about the infringement of their intellectual property and satisfied the funding requirements in data publication [7, 29, 30, 32].

2.1 Project Structure

The first step involved developing a domain-specific ontology for materials science that was capable of integrating the semantic information of the heterogeneous data from disparate, but relevant, databases and modelling the entire provenance of the compound composition program. The ontology that was developed is described in Chapter 3.

The second step involved prototyping a federated search interface based on the materials science ontology as a single Web-based search interface to the high priority materials science databases identified by our collaborating scientists. This enabled scientists, within a single platform, to search, retrieve, correlate and integrate diverse, but related, data for further analysis. The search interface that was developed is described in Chapter 4.

The third step involved investigating data mining and advanced atomistic modelling techniques for the discovery of knowledge from large volumes of multivariate data and the simulation of complex oxide systems, respectively. There will be two related sub-projects, the first will develop an interactive data mining with visualization to identify patterns that are common within the parameters of those compounds identified by the tool developed in the previous step. The other will apply well-established energy minimisation methods to accurately modelling defects, local structures and ion migration mechanisms in complex oxides for potential Solid Oxide Fuel Cells use. These will provide sets of high-quality experimental parameters for the compound synthesis program. However, these projects are outside the scope of this thesis.

The fourth step involved prototyping a workflow management system that, (1) generated the workflow instance to manage the compound synthesis program, (2) coordinated, orchestrated and interoperated the collaborating scientists distributed geographically and self-contained computational services, (3) automated and monitored the execution of the workflow instance and, (4) precisely captured and recorded the data and metadata of every phase of the experiment workflow. This provided quality control for the experimental process insured against data loss and ensured the captured data would be in good quality and consistency. This prototype system is detailed in Section 2.5.

The last, the fourth, step involved developing a prototype system that rendered the intuitive, customized graphical view of the scientific experiment workflow that was generated and executed by the system described in Section 2.5. The prototype also enabled scientists to, (1) select data nodes from the workflow view and import discoverable objects via the Web into a scientific compound object [56], (2) infer the relationships between the individual components within the compound object through semantic inferencing rules and a rule-inference engine, (3) attach a copyright license to compound object and, (4) comply with the Open Archives Initiative Protocol – Object Reuse and Exchange (OAI-ORE) [128], which can be saved in standardised web-publishing formats. The work that was developed is described in Chapter 5.

For the full capability of the data-driven approach to be realized, researchers wanted a workbench that was able to satisfy the following requirements. Initially, it would enable the users to search and assimilate diverse, but related, materials data through sophisticated semantic search and integration techniques. The mapping of the semantic information of the data was developed and evaluated using Semantic Web technologies, including OWL and SPARQL. Next, in addition to the data mining and atomistic modelling components that are outside the scope of this thesis, the workbench incorporated a workflow management system that generated, automated and monitored the workflow instance to manage the compound synthesis program. Additionally, the system also captured and recorded the data relating to the workflow instance systematically to ensure the data had high-quality and consistency. The workflow approach was developed using the Service-oriented Architecture and WSBPEL technologies, including Apache Axis and IBM's BPWS4J. Finally, the workbench also incorporated a component that was able to overcome some of the current barriers to scientific data publication that include, (1) a lack of incentive, (2) a lack of tools, (3) difficulty in preparing data for publication, (4) difficulty providing the appropriate level of provenance data and, (5) concern with intellectual property rights. This functionality was developed and evaluated using Semantic Web and OAI-ORE technologies, including OWL, Semantic Web Rule Language (SWRL), and their APIs, such as JENA and Protégé-OWL, Algernon, the rule based engine and FORESITE's APIs.

These techniques were developed using the fuel cell domain described in Section 1.3 as a test bed. One aim was to test the theory that, (1) mapping heterogeneous data through a domain-specific ontology, (2) managing the scientific process through workflows, (3) capturing data generated during the workflow execution and, (4) publishing scientific datasets complying with OAI-ORE would all work better in narrow, deep domains, such as the fuel cell domain.

The remainder of this chapter describes the framework of the proposed Materials informatics Workbench (using the fuel cell domain as an example application) and presents the tasks in data integration, workflow management and data publication, covered by the workbench.

2.2 Architectural Overview of Workbench Framework

Figure 2.1 illustrates the purpose of the workbench framework using the domain described in Section 1.3. The architecture is loosely coupled and allows more subsequent functionality to be plugged in, such as the data mining and atomistic modelling components.

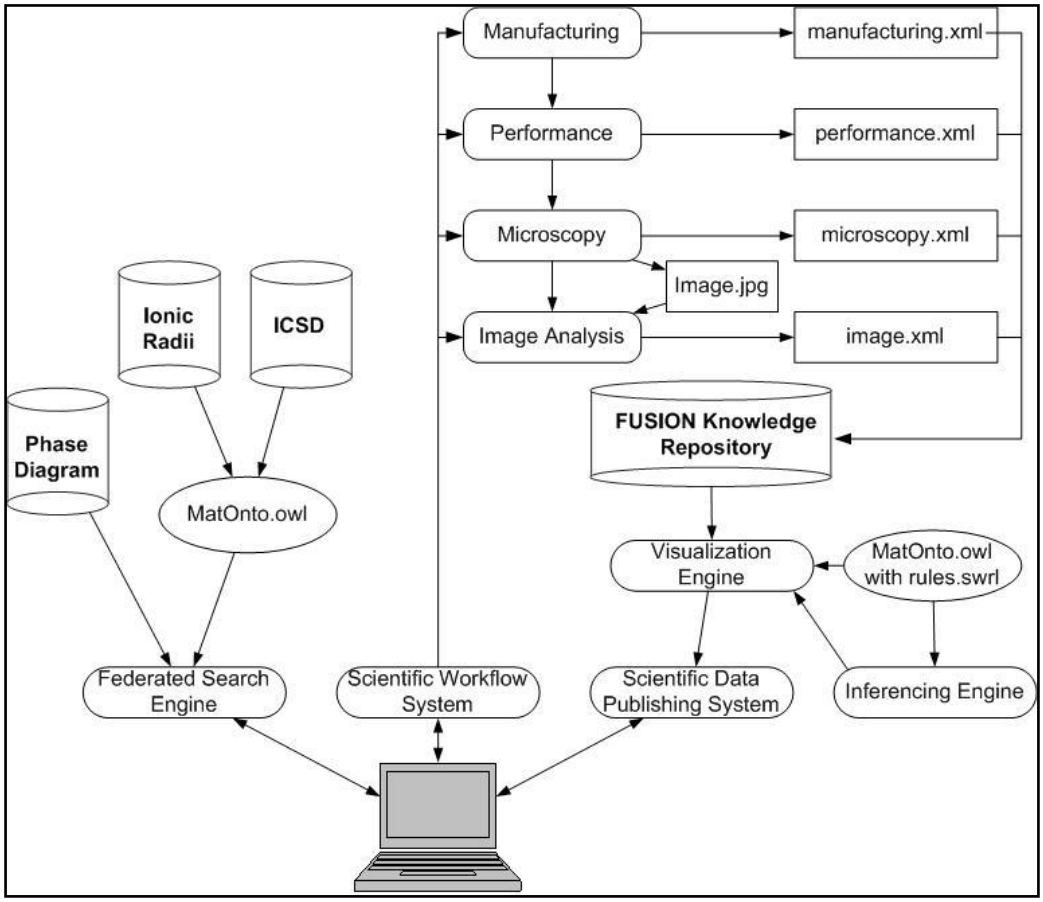


Figure 2.1: Framework Overview

There are three key tasks approached by this framework:

- a search interface for searching and integrating diverse data through a domain-specific ontology
- management of the scientific investigation process through workflows and data capture
- a user interface for data access, author and publication.

This workbench framework enables materials scientists to, (1) search, retrieve and integrate a wide range of data from disparate databases through a single platform, (2) manage the scientific investigation process through the workflow system, capture and record experimental data and metadata precisely, systematically and, (3) intuitively access, author and publish scientific data sets

complying with OAI-ORE. This maximises the potential of the data-driven approach for expediting the design and discovery of novel compound materials.

The following sections walk through the workbench framework using the case study described in Section 1.3 to describe each of these key tasks.

2.3 Ontology

The initial step in any data management project spanning across relational databases involved understanding the semantic difference between metadata terms from the database schemas and developing a data model that resolves the difference. This is particularly important and challenging when the aim is to support the advanced querying, correlation and integration of large volumes of heterogeneous data. To do this, a semantic layer that provides a formalized model is required. Because of the necessity of the exchange, re-use and integration of materials science data and experimentation, a domain-specific ontology for materials science was used as a common, extensible data model. The ontology represents structured knowledge about materials, their structure and properties and the processing steps involved in their composition and engineering.

First, the ontology enables the integration of the heterogeneous data from the designated materials science databases. Second, the ontology is used to model the provenance information captured by the workflow management system, thereby facilitating the subsequent data access. Finally, the ontology combined with the SWRL rules enables the inferencing through a rule-reasoning engine of the coarse-grained view of the scientific methodology from the fine-grained provenance trails of a complex scientific workflow.

Figure 2.2 highlights the materials science ontology. The implementation approach taken will be expanded upon in Chapter 3.

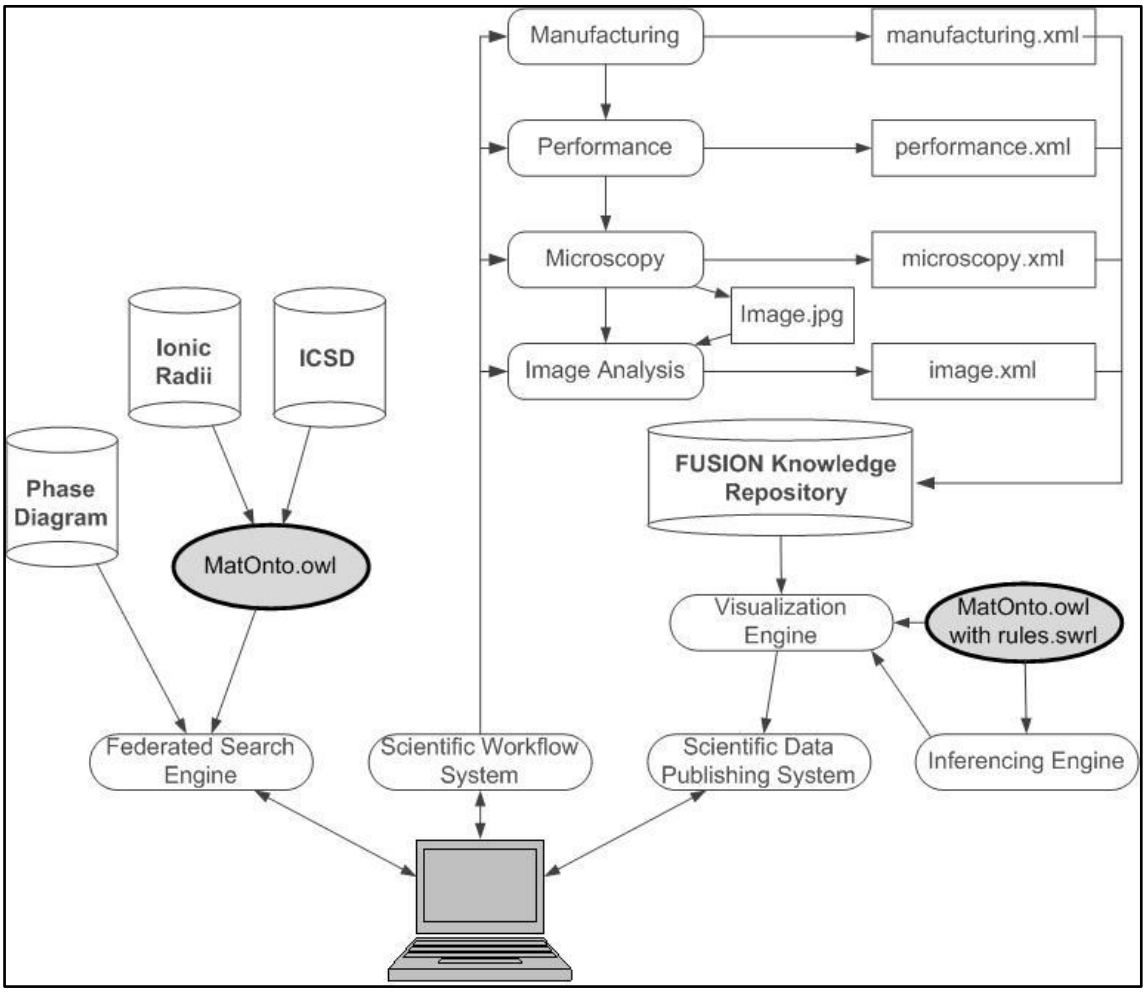


Figure 2.2: Materials Science Ontology

2.4 Data Integration

It is clear that generating a very precise and targeted compound synthesis program required statistically analysing large volumes of disparate but relevant scientific data. Manually searching, retrieving, correlating and integrating heterogeneous data across autonomous databases reduces the materials scientists' efficacy greatly for collecting and correlating the precise data sets because of the difficulty of resolving the differences in the semantics of the metadata terms, data structures, formats and metrics. Therefore, part of the workbench framework developed for this project was focused on the methods and tools for the fully-automated resolution of the differences in the semantic information between the heterogeneous data, and correlating and integrating the data seamlessly. The development of the materials science ontology is the first step towards solving this issue. Based on the ontology, a federated search interface to the designated materials databases was developed.

Materials databases identified by our collaborating scientists include the Inorganic Crystal Structure Database (ICSD), Ionic Radii Database and the NIST Phase Equilibria Diagrams Database. The materials science ontology underlying the search interface provides users with search keywords,

while it is also a means for resolving semantic inconsistencies by mapping ontological terms to metadata terms from the involving database schemas; thus, the users can seamlessly interrogate a wide range of data across disparate databases. Furthermore, the search interface is also a single entry point for useful tools, including rendering 3D crystal structures, calculating bonds length and angles, exporting Crystallographic information files and retrieving relevant scholarly papers. Therefore, through the search interface, the materials scientists can use familiar keywords for searching and retrieving answers to queries such as, 'Give me compounds that contain tungsten-oxygen-X (where X is a different cation), with bond lengths between Y and Z nm, with large anomalies and anisotropy in the optional parameters of oxygen, with bond angles between J° and K° and which are stable below 500°C'.

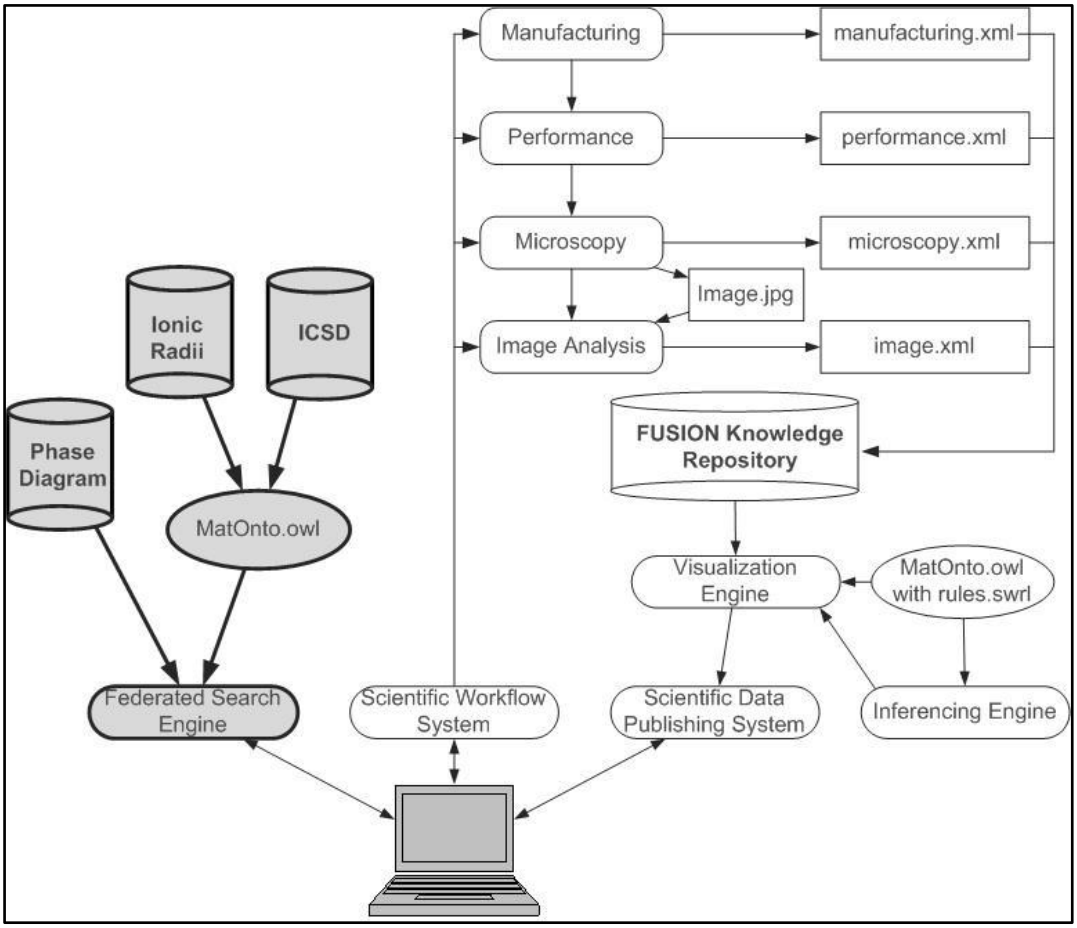


Figure 2.3: Data Integration Components of the Workbench Framework

Figure 2.3 highlights the components of the federated search interface, underlying ontology and involving databases within the workbench framework. The search interface is a combination of Semantic Web and Web 2.0. The former provides search keywords and enables the resolving of the inconsistent semantic information between databases through OWL and SPARQL. The latter

provides a better human-computer interaction through the AJAX libraries. The technologies applied and the implementation approach taken will be expanded in Chapter 4.

2.5 Managing Workflows and Data Capture

After generating a very precise and targeted compound synthesis program through data analysis and atomistic modelling to be developed, the program is transformed into the workflow instance and managed by a workflow management system [129, 130] underpinning a workflow engine complying with Web Service Business Process Execution Language (WSBPEL) [54]. The system enables the coordination, orchestration and interoperation between self-contained computational services and cross-disciplinary scientists who are across multiple organizations and are in charge of different experimental activities within the compound synthesis program. Additionally, it also captures and records all data generated by the workflow instance accurately and systematically to ensure the data is high-quality and consistent. This also lays a good base for subsequent data access.

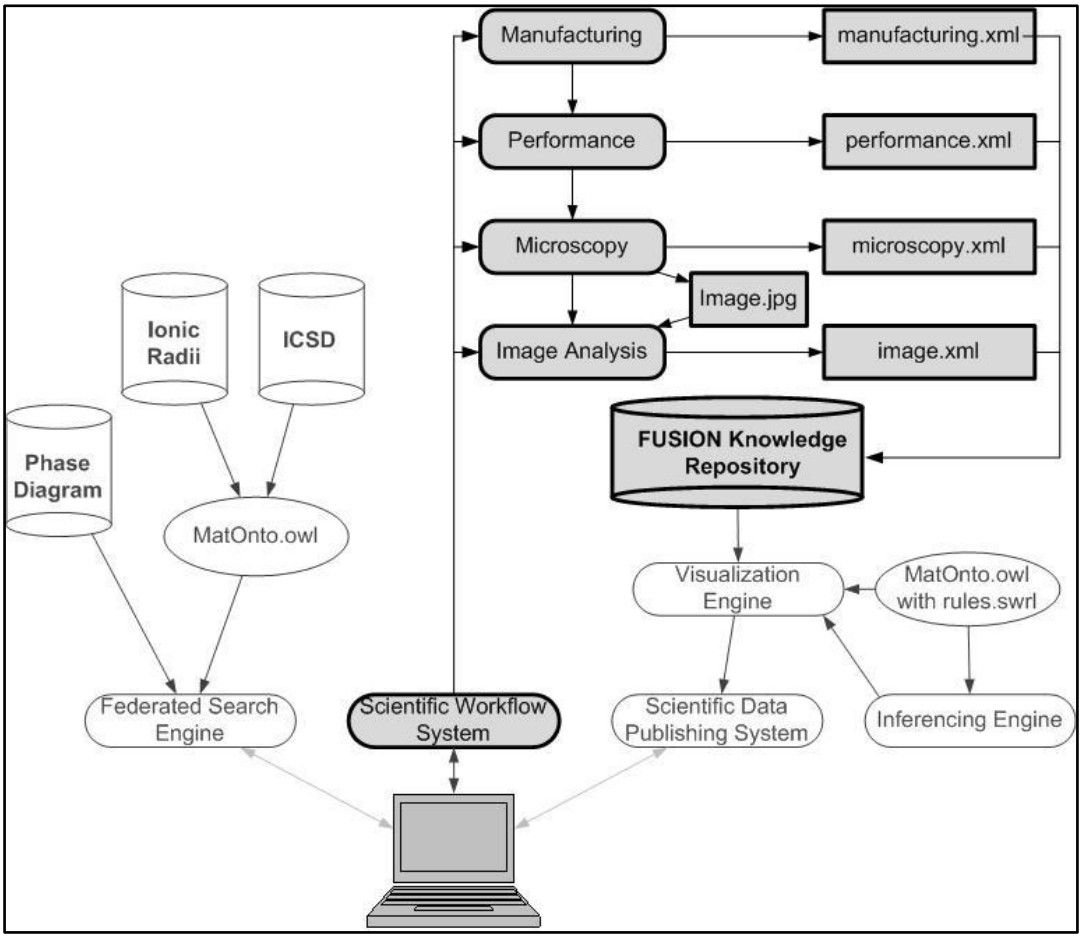


Figure 2.4: Workflow and Data Capture Components of the Workbench Framework

Figure 2.4 highlights the components of the workflow management within the workbench framework. The following briefly describes the system in the aspects of system architecture and functionality through a case study.

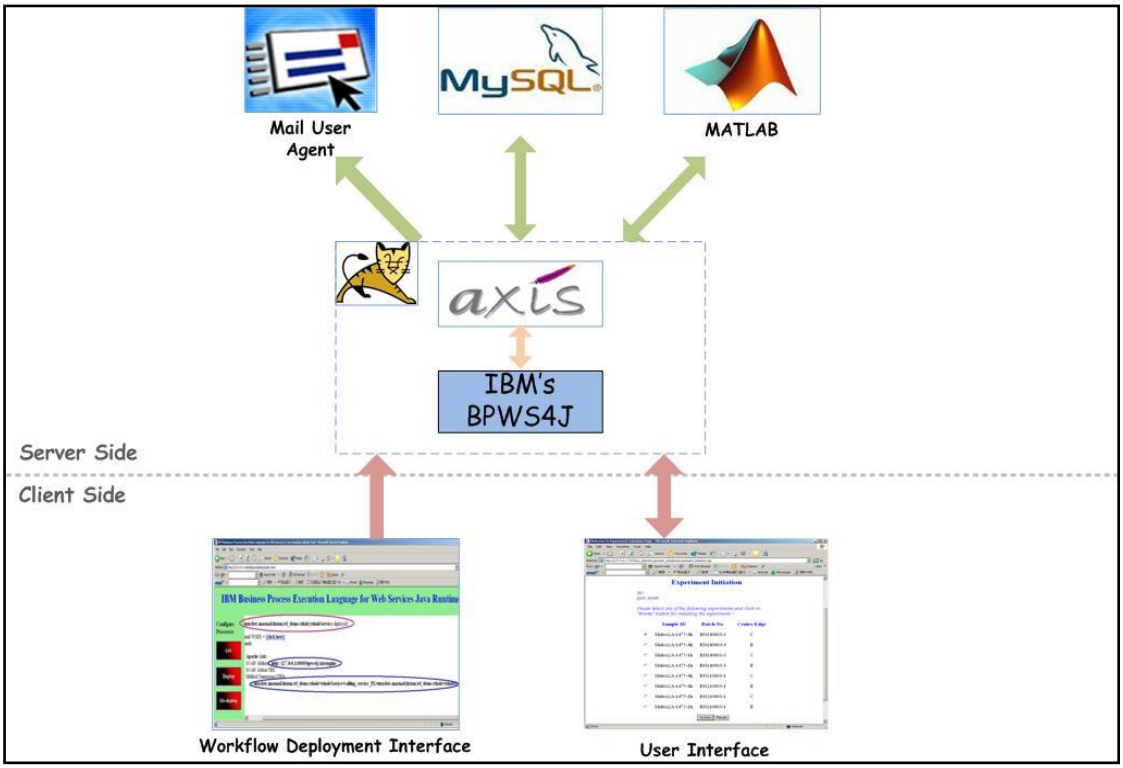


Figure 2.5: Apache Tomcat-based Web Services Architecture

2.5.1 System Architecture

Figure 2.5 illustrates the overall architecture and major components of the WSBPEL workflow management system. There are four major components, the Apache Tomcat [131], MySQL Database Server [132], MATLAB [133] and the Mail User Agent [134]. First, the Apache Tomcat server hosts the Apache Axis — an SOAP engine [135]. Four critical web services have been deployed to Apache Axis for this system. One of them is the IBM Business Process Execution Language for Web Services Java Run Time (BPWS4J) — a core engine of this system [136]. The next is a web service for the access to one of MATLAB's image analysis functions through which the BPWS4J sends a microstructure image to MATLAB and captures the analytical results from the image analysis. The third is used for the access to the MySQL Database Server that holds the scientific workflow information about equipment settings and instrument operators. The last is an alerting web service to notify participating instrument operators by sending them emails when their tasks are available.

2.5.2 The Manufacture of Oxygen Ion Conductors – an Example Scenario

The manufacturing process for oxygen ion conductors is extremely complex and we will not attempt to describe it here. Figure 2.6 merely illustrates the steps involved in the manufacturing and testing process. All of the tasks are human operations except for the Microstructure task, which is conducted by MATLAB. The corresponding workflows in the Business Process Modelling Notation BPMN are in Appendix B including the high level view of the overall process and the lower level views of the human-and computational-activity sub-processes.

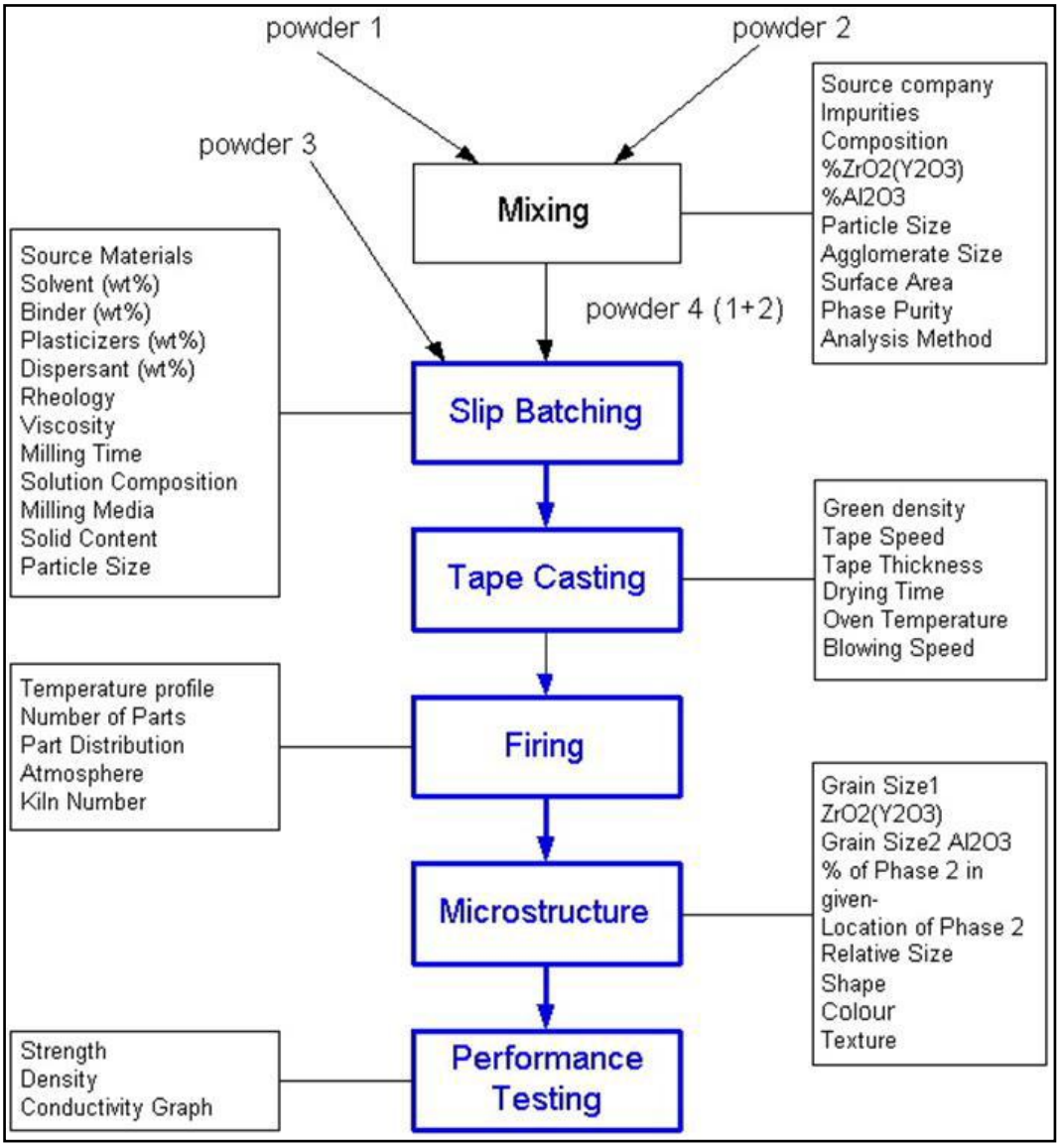


Figure 2.6: Manufacturing and Analysis Workflow for an Oxygen Ion Conductor

2.5.3 System Functionality

When an authorized user logs on to the system, they are presented with a web page for experiment initiation. Figure 2.7 demonstrates a number of experiments with different settings being available. When the user selects one of them, a page for the task allocation is rendered. Figure 2.8 demonstrates that there are four drop-down boxes, including a group of names of potential operators across different organizations available for the tasks of Slip Batching, Tape Casting, Firing and Performance Testing. Then, the BPWS4J engine is invoked and automatically executes the selected experiment workflow instance.

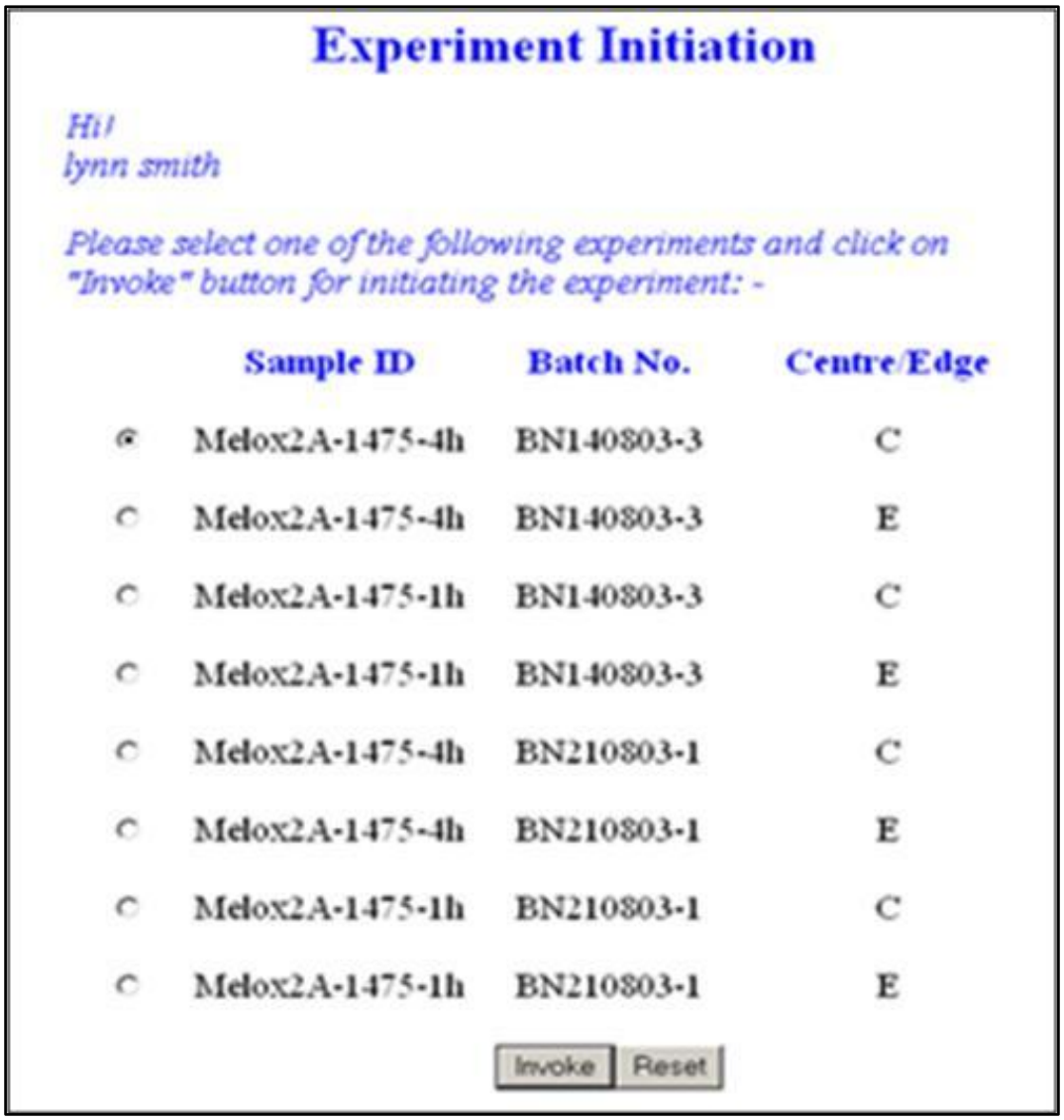


Figure 2.7: The Web Interface for Experiment Initiation

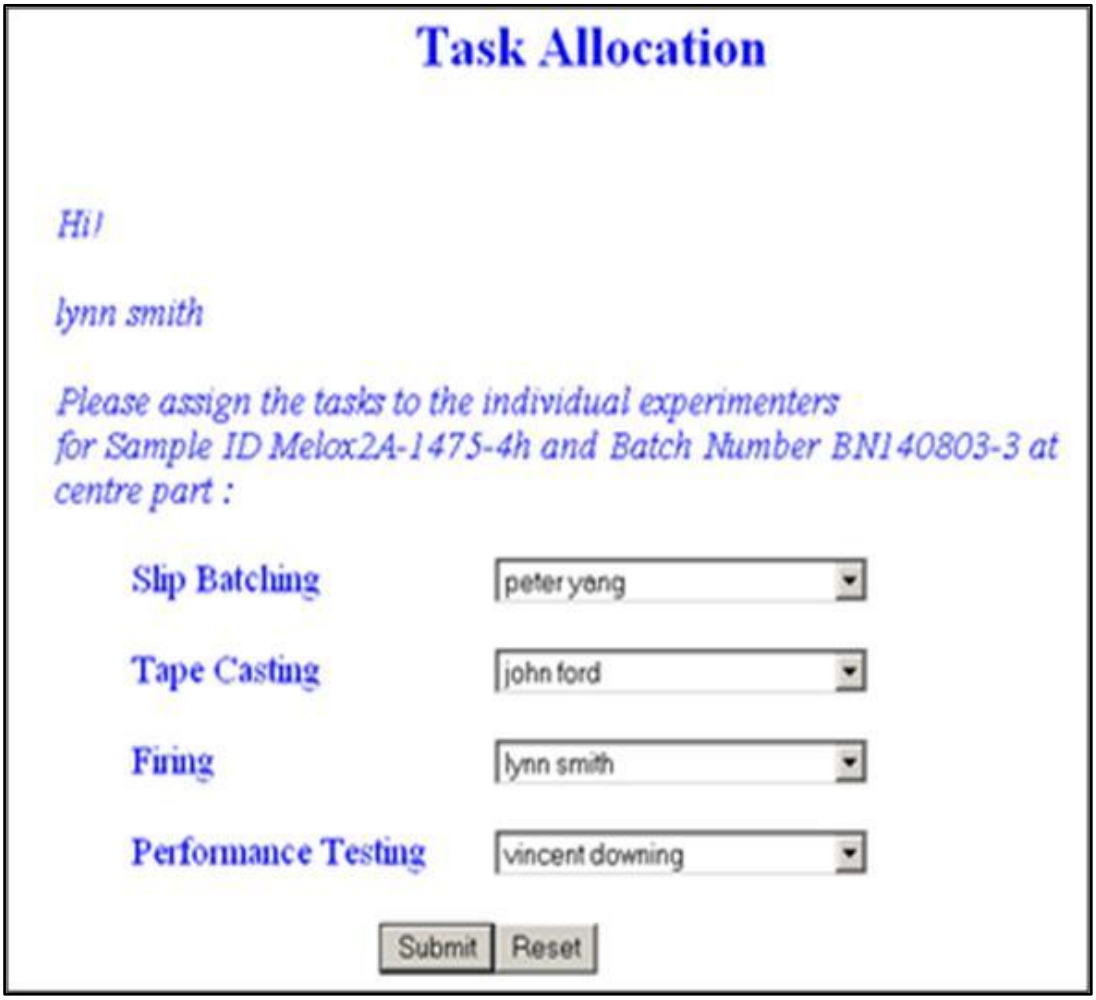


Figure 2.8: The Web Interface for Task Allocation

As the workflow advances, the engine notifies the participating operators by sending them email through the invocation of the alerting web service when the task is available. The email includes a URL for the parameters of the equipment settings. When the operator accesses the parameters through the URL, the system invokes the database access web service to retrieve the data and renders a web page that displays the data. Figure 2.9 demonstrates a web page displaying the parameters for the operator to setup the relevant equipment in the Slip Batching task.

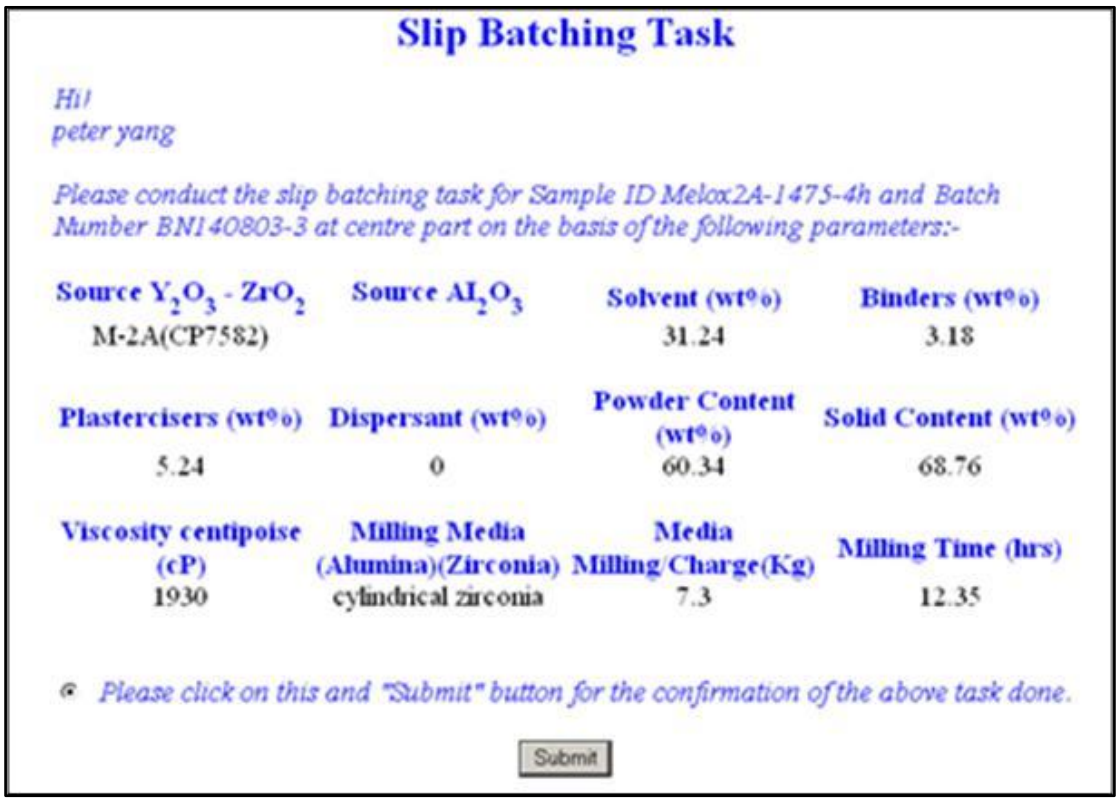


Figure 2.9: The Web Interface for the Equipment Settings

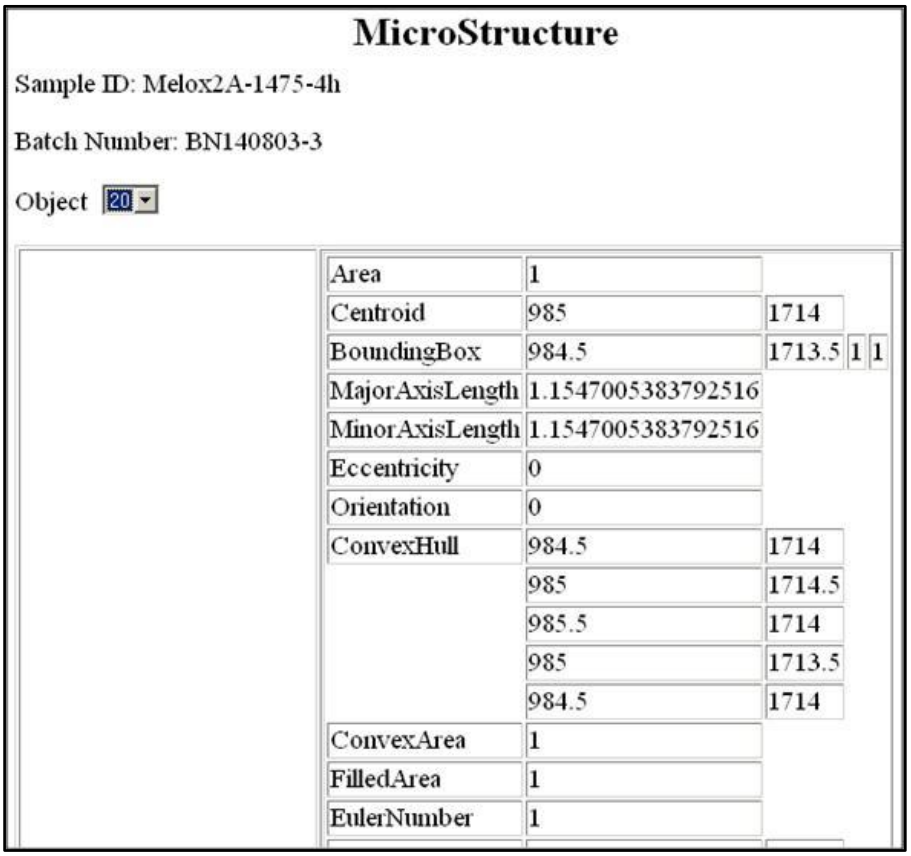


Figure 2.10: The Web Interface for the Analytical Results of Image Analysis

In addition to the human activities, including Slip Batching, Tape Casting, Firing and Performance Testing, there is a computational task — Microstructure Analysis — that is conducted through the invocation of a web service interfacing one of image analysis functionalities within MATLAB. When this task is executed, the system, (1) invokes the web service with the URL of the image resulted from the Characterization process as an input, (2) captures the multivariate and multidimensional numerical data as the analytical output generated by the web service and, (3) stores the data into a designated distributed database system. When a user requires a reading of the data, the system invokes the database access web service for data retrieval and renders a page displaying the data shown in Figure 2.10. Finally, a user is able to monitor the progress of the runtime workflow through a web interface shown in Figure 2.11.

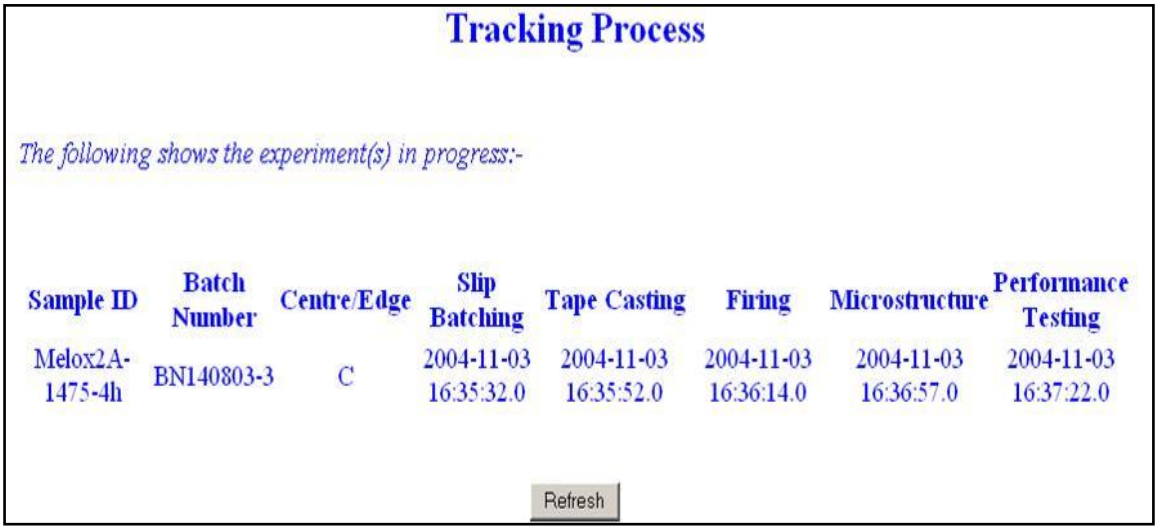


Figure 2.11: Workflow Monitor User Interface

2.5.4 User Feedback

Following the trials, our collaborating AIBN scientists felt the surge of relief that the complex, cross-organization compound synthesis program could be managed by the WSBPEL workflow system that incorporates the human and computational activities. They were also impressed with its ability for capturing and storing huge amounts of multivariate and multidimensional analytical data generated by the computational service that are no longer tackled manually with traditional paper-based laboratory notebooks. This systematic approach to data collection facilitates the subsequent data access. Amazingly, they also realised that a scientific workflow with sufficient provenance information encapsulates scientific intellectual property and enables the sharing of knowledge between scientists, for example, via the myExperiment initiative [137].

2.6 Data Access, Authoring and Publication

The systematic capture and recording of experimental data during the execution of the workflow instances facilitates the dissemination of scientifically valuable datasets. Rendering an intuitive graphical view of the scientific investigation process for scientists is an ideal way of accessing recorded data. The visualized scientific process consists of nodes and arrows. The former represents either experimental data or activities, while the latter indicates the relationships between the nodes. Scientists can easily access the information of required data through the data node with the associated metadata. Additionally, this mechanism also lays the base for the publication of scientific data.

Publishing scientific datasets with contextual/provenance information is essential for validation and verification, and the reproduction of the research results. Using the SWRL rules with a rule-reasoning engine enables the scientist to infer the coarse-grained view of the scientific methodology from the fine-grained view of the complex scientific workflow provenance trail. From the digital librarian perspective, a package of scientific datasets encapsulating the scientific methodology with critical datasets is equivalent to a scientific compound object [56]. A scientific compound object enables scientists to encapsulate the various datasets and resources generated or used during a scientific experiment or discovery process, within a single compound object, for publishing and exchange. Furthermore, the compound object, compliant with OAI-ORE in standardized web publishing formats, becomes machine-readable, discoverable, exchangeable and reusable.

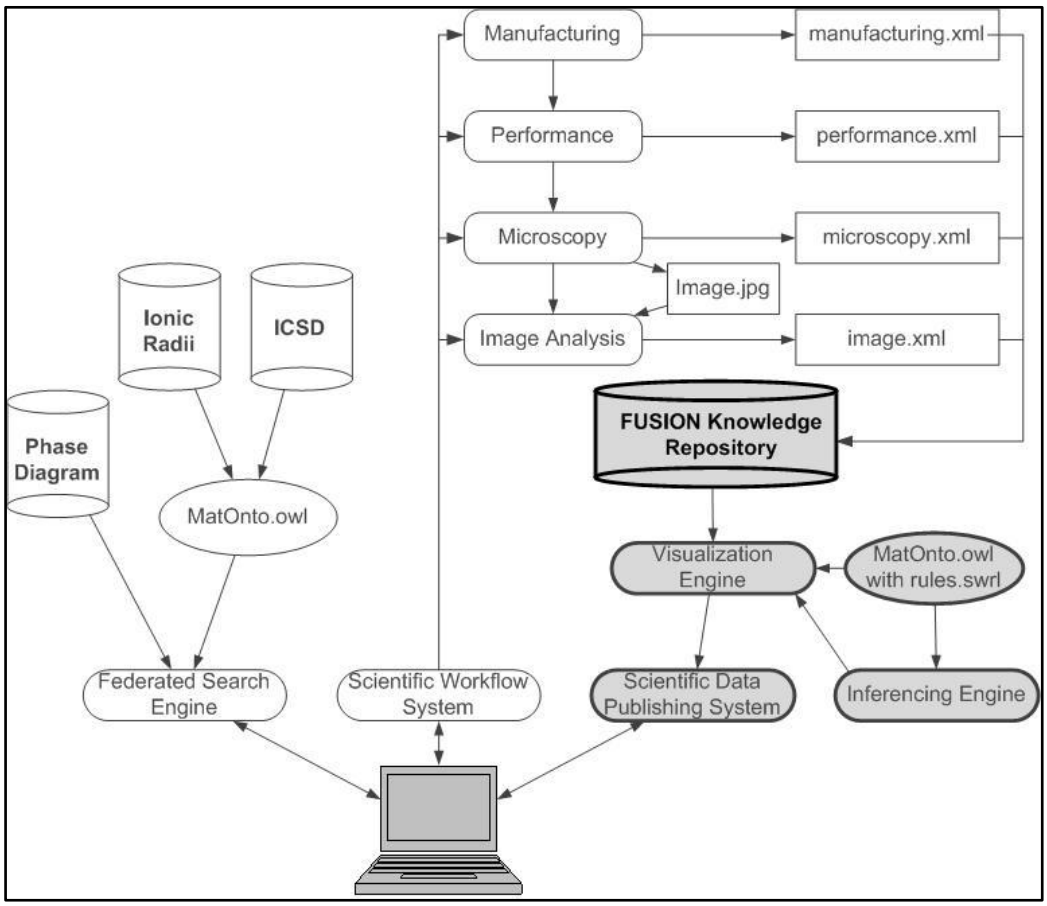


Figure 2.12: Data Access, Authoring and Publication Components

Figure 2.12 highlights the components of a scientific data publication that provides a simple way for scientists to access acquired data and to author and publish scientific data. Chapter 5 describes in detail the technologies and experiences of implementing the prototype system.

2.7 Summary

The aim of this thesis is to apply, evaluate and extend the emerging Semantic Web, WSBPEL and OAI-ORE technologies to provide innovative approaches for semantically integrating materials science data and information, managing the scientific process through workflows and capturing data, and publishing scientific data sets in standardised formats. This chapter has outlined the workbench framework that can support these requirements by:

1. Using a domain-specific ontology to enable the semantic information integration of different types of data
2. Generating, automating and monitoring the scientific experiment workflow
3. Supporting the capture of high-quality, fine-grained, structured data, metadata and provenance information
4. Applying semantic inferencing rules for developing a coarse-grained view of the scientific methodology from the fined-grained view of the complex scientific workflow provenance trails
5. Providing innovative interfaces for accessing, authoring and publishing scientific data complying with OAI-ORE in the machine-readable, discoverable, exchangeable and reusable formats.

The next chapter expands the approach presented in Section 2.3 and describes the Materials Science Ontology that underpins the federated search interface and scientific data publication components discussed in Sections 2.4 and 2.6, respectively, and the process through which the ontology was developed.

Chapter 3 MatOnto – Materials Science Ontology

3.0 Introduction

This chapter presents the Materials Science Ontology (MatOnto) that is an underlying semantic foundation for, (1) data integration through a federated search interface to the key materials science databases described in Chapter 4 and, (2) scientific data publication through a scientific compound object publishing system described in Chapter 5. The MatOnto is an extensible framework that encapsulates the top level structured knowledge of materials science to enable:

- mapping between and integration of disparate databases within the materials science domain
- modelling and capture of precise provenance data generated by the scientific experiment workflow — this is essential to enable verification, validation, comparison and reuse of experimental results
- inferencing and extraction of new knowledge in the materials science domain, through the application of SWRL rules and a rule-reasoning engine.

This chapter is structured as follows:

- Section 3.1 describes MatOnto's design philosophy, critical modules and development process
- Section 3.2 assesses MatOnto through Gruber's five design criteria for ontologies
- Section 3.3 describes MatOnto's evaluation, limitations and future work.

3.1 Development

MatOnto's design principles are to provide an ontology that:

- is based on an upper ontology, an advanced knowledge representation system that is a library of richly structured and well-understood abstract data types and structural organizational principles that make the technical aspects of ontology construction easier and more reliable [138]
- leverages existing peer-reviewed ontologies or vocabularies developed through community consensus
- enables the integration of those high priority databases identified by the collaborating scientists.

Below we describe the six steps in the process of developing the MatOnto ontology.

First, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [139], the upper ontology developed by the Laboratory for Applied Ontology (LOA), is selected as the upper basis for MatOnto. As reflected by its name, DOLCE has a clear cognitive bias and aims at capturing the ontological categories underlying the natural language and human commonsense. The categories are merely descriptive notions that assist in making the already-formed conceptualization explicit. DOLCE stems from the *Particular* root class. *Particular* has four subclasses: E*ndurant, Perdurant, Abstract and Quality*, from which we defined the MatOnto subclasses.

Second, a number of existing peer-reviewed ontologies and a classification system were leveraged, Ontolingua's Standard Units and Dimensions [140, 141], the Joint Academic Classification of Subjects (JACS) [142], The World Wide Web Consortium (W3C)'s Time Ontology in OWL [143] and, AIFB's Semantic Web for Research Communities (SWRC) ontology [144].

- Ontolingua [141] is a mechanism that is used to represent Gruber and Olsen's Engineering Mathematics Ontology [140]. It includes conceptual foundations for scalar, vector and tensor quantities, physical dimensions, units of measure, functions of quantities and dimensionless quantities. MatOnto incorporates two of them — units of measure and physical dimension.
- The JACS system is used to classify academic subjects by the Higher Education Statistics Agency [145] and the Universities and Colleges Admissions Service [146] in the United Kingdom.
- W3C's Time Ontology in OWL describes the temporal content of web pages and the temporal properties of web services. The ontology provides a vocabulary for expressing facts

about topological relations among instants and intervals, together with information about durations and about date-time information.

- The SWRC ontology models research communities and relevant related concepts.

Third, I extended a common ontology of scientific experiments (EXPO) [147], an ontology for describing scientific experiments with the ABC Metadata Ontology [148], to enrich the EXPO with the concepts of *events* and *processes*. EXPO is not only a taxonomy of scientific experiments, but also aims to be the formal description of scientific experiments for efficient analysis, annotation and the sharing of scientific results. On the other hand, the ABC Ontology models events in both the physical domain and a digital object's lifecycle.

Fourth, I developed the top-level ontology for materials science according to the materials science textbook [149] and the Springer Handbook of Materials Measurement Methods [150], beginning with the class *matonto:Material* that is linked to *jacs:MaterialsScience* of the JASC system.

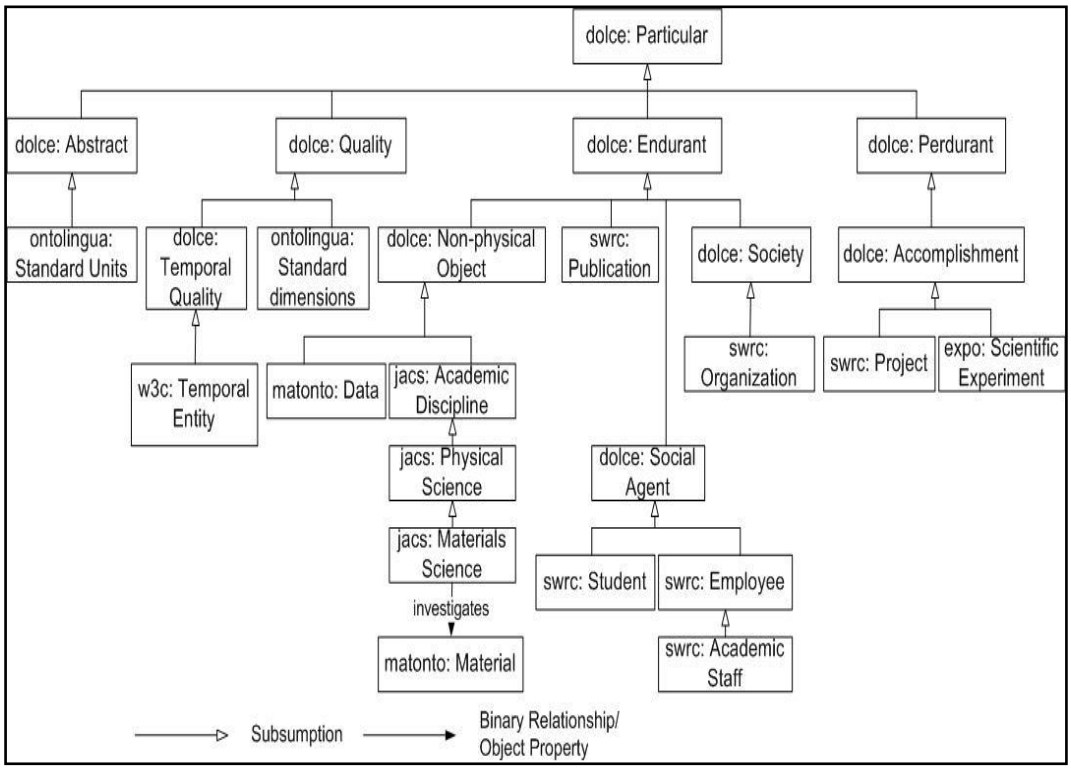


Figure 3.1: The Top Level Classes

Figure 3.1 represents the complete top-level view of MatOnto — the classes with the prefix *dolce* are from DOLCE. Figure 3.1 shows the use of classes (with prefixes *ontolingua*, *swrc* and *w3c*, respectively) from the existing peer reviewed ontologies[140, 141, 144] . It illustrates the use of classes from the JACS system [142] (those with prefix *jacs*), the EXPO ontology [147] (prefix *expo*) and the root class of our MatOnto ontology — *matonto:Material*.

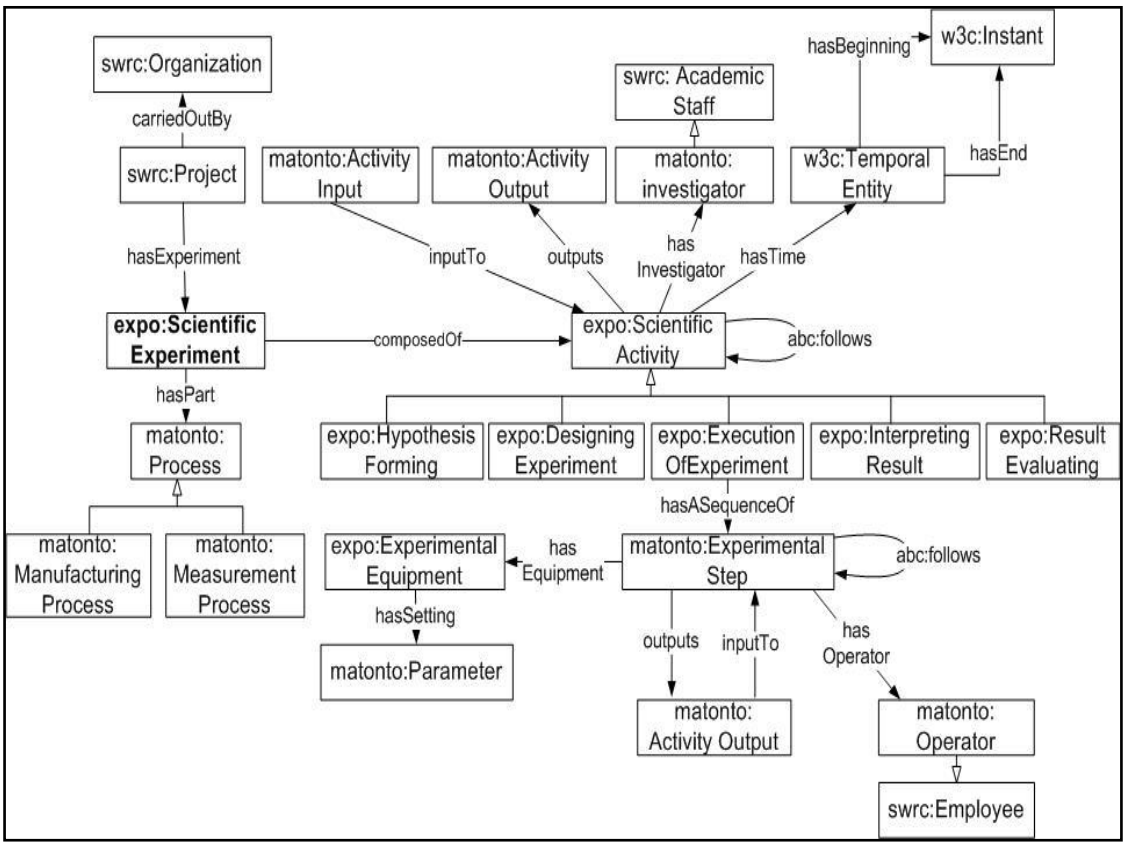


Figure 3.2: The Merging of the EXPO and ABC Ontologies

Figure 3.2 demonstrates the merging of the classes from the EXPO and ABC ontologies that model the event-oriented scientific experiments.

Figure 3.3 shows five core properties associated with *matonto:Material*:

1. *matonto*:*Property* — the materials properties
2. *matonto:Family* — the materials classification
3. *matonto*:*Process* — the materials manufacturing and measurement processes
4. *matonto*:*Structure* —the materials structure
5. *matonto:MeasurementData* — the data resulting from the measurement or the characterisation process. We have drilled down to certain levels and structured the associated concepts in a logical way.

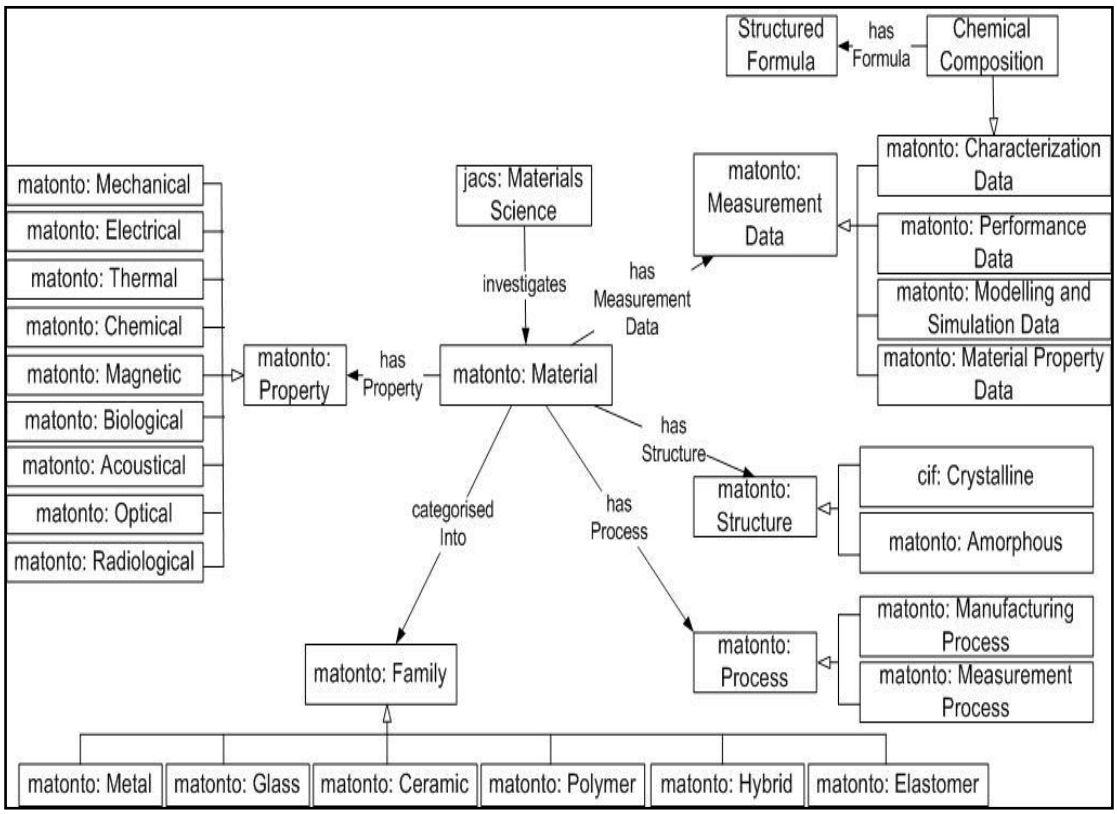


Figure 3.3: Materials Properties, Family, Processes, Structures and Measurement Data

Fifth, we developed a sub-disciplinary ontology describing the concepts associated with crystalline structures according to the Crystallographic Information Framework [151] and started it with the class *cif:Crystalline,* which is a sub-class of *matonto:Structure*. Figure 3.4 demonstrates the complete top-level view of the Crystalline Structure Ontology.

Finally, we developed a simple scientific data ontology by incorporating the classes from the Suggested Upper Merged Ontology (SUMO) [152] and the MPEG-7 ontology [153] for describing the numerical and multimedia data, respectively. Figure 3.5 illustrates the high-level view of the ontology.

Linking of all of these sub-ontologies via their common classes generates the complete MatOnto ontology.

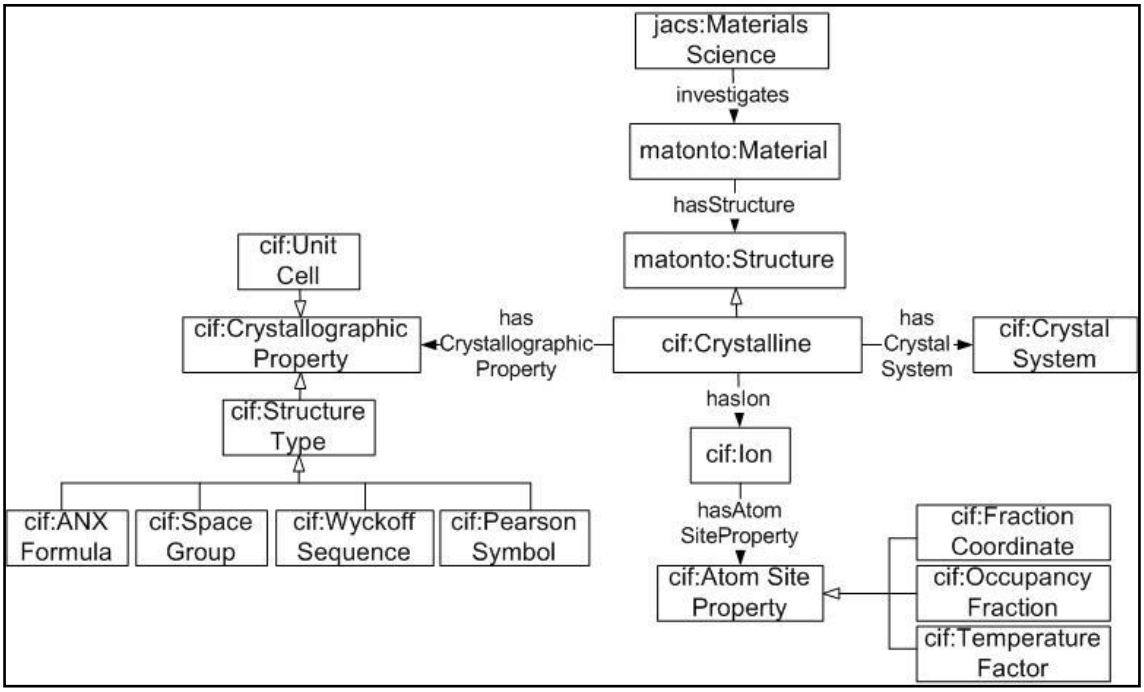


Figure 3.4: Crystalline Structure Ontology

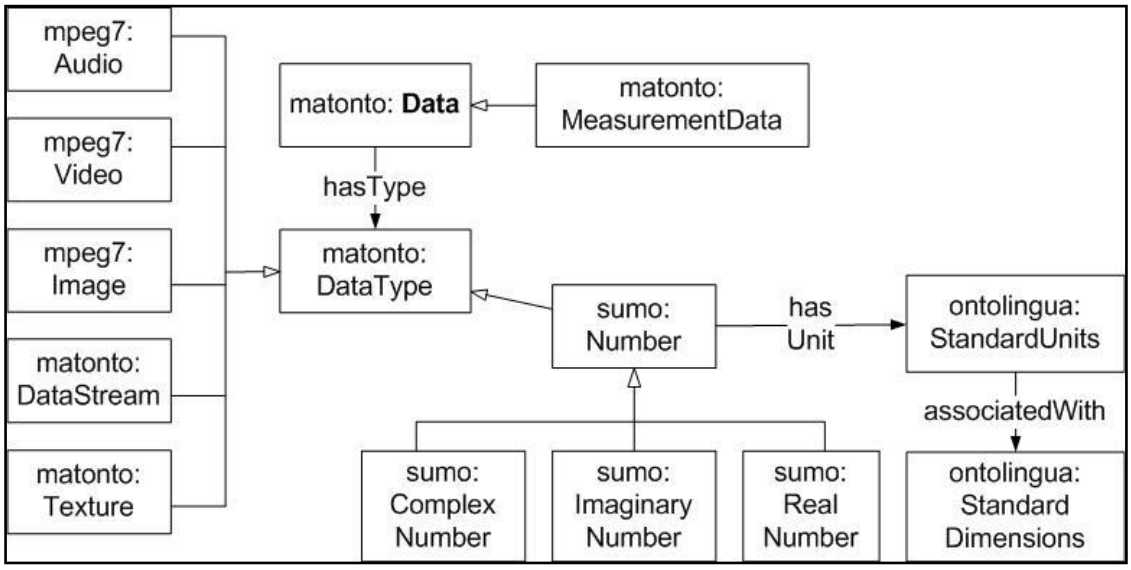


Figure 3.5: Simple Scientific Data Ontology

3.2 Assessment

MatOnto's quality has been assessed based on Gruber's five criteria [77] — clarity, coherence, extendibility, minimal encoding bias and minimal ontological commitment — with satisfactory results. First, MatOnto possesses clarity, because its vocabulary is sourced from peer-reviewed ontologies and existing standardized taxonomies. Second, MatOnto does not have incoherency issues, because no concepts are derived via inferencing. Third, MatOnto is extensible, because DOLCE together with JACS provides a proven platform for integrating disciplinary ontologies. The high-level materials science ontology provides a platform for integrating sub-disciplinary ontologies within the materials science domain, for example, the Crystalline Structure Ontology. Fourth, MatOnto has no encoding bias, because it is free of implementation details. Finally, MatOnto has a low ontological commitment, because we have reused existing peer-reviewed ontologies and extended them, based on standardized vocabularies.

3.3 Evaluation

MatOnto satisfies the objectives outlined in the beginning of this chapter. It enables the integration of the existing related sub-disciplinary and relevant ontologies through the top-level materials science classes shown in Figure 3.1. The Crystalline Structure Ontology shown in Figure 3.4 enables the integration of and the mapping [154] between the Inorganic Crystal Structure Database (ICSD) [1] and the Ionic Radii database [155]. The details are expanded in Chapter 4. The extended EXPO ontology shown in Figure 3.2 enables the capture of precise provenance data and the inferencing of new knowledge (for instance, relationships between nodes that are not explicitly related). This aspect is used to automatically infer coarse- grained views of the scientific methodology from a fine-grained view of the complex scientific workflow provenance trails for publication or e-learning purposes [156]. The details are discussed in Chapter 5.

3.4 Limitations and Future Work

Currently, this ontology has just been used by the AIBN scientists. Additionally, it has been submitted and reviewed by the materials science community. The initial feedback is positive [157]. Hopefully, engaging the community members in the future evolution will help reach a greater consensus within the community.

We have a plan with the collaborating scientists to develop a set of SWRL rules to infer new implicit relationships and knowledge from explicit data in a number of aspects of materials science, including:

- Inferring relationships between processing parameters and structure
- Inferring relationships between structure and properties or behaviour
- Inferring structural features from automatic image analysis of microscopy images.

3.5 Summary

In this chapter, we have described MatOnto — an ontological framework that encapsulates the knowledge structure of materials science and that can be easily extended to integrate with related ontologies. MatOnto enables the materials scientists to search, retrieve and integrate data from heterogeneous and disparate data sources, based on a common set of ontological terms detailed in Chapter 4. It also enables the capture of the processing steps and provenance information both within the laboratory and within the computing environment. This enables the repeatability, exchange, comparison and re-use of the experimental results detailed in Chapter 5. The MatOnto ontology also provides the potential for inferencing and extraction of new knowledge using SWRL rules defined by domain experts (for example, fuel cell scientists) and a reasoning engine (Pellet). MatOnto is an integral component of the cyber-infrastructure for the materials science community as discussed in Section 1.1.1. The next chapter will describe a federated search interface underpinning MatOnto for the key materials science databases and analytical tools. The MatOnto in the Manchester OWL is in Appendix A.

Chapter 4 MatSeek – Ontology-based Federated Search Interface

4.0 Introduction

This chapter presents the MatSeek system [154] — an ontology-based federated search interface to key materials science databases and analytical tools. By combing Semantic Web and Web 2.0 technologies, MatSeek provides materials scientists with a single Web interface that enables them to, (1) search across disparate databases containing crystal structure data, ionic conductivity data and phase stability data, (2) to render 3D crystal structure images, calculate bond lengths and angles, (3) retrieve relevant scholarly references and, (4) identify potential new materials with the structure and properties to satisfy specific applications. MatOnto underpinning MatSeek enables the mapping between the databases schemas and the consequent integration of the data. It also enables the construction of query statements dynamically and accurately, thereby resulting in an intuitive, Google-like, user-friendly search interface. The Web 2.0 technologies enable iterative searching across the databases — the retrieved results from searching the previous database are used as input to the query on the next databases.

The remainder of this chapter is structured as the follow:

- Section 4.1 describes the system architecture including the server and client sides and the interaction between them
- Section 4.2 abstracts the technical details including the MatOnto ontology, Referential Relationship ontology, manual mapping of database schemas through MatOnto, dynamic construction of SQL query statements and data correlation and integration
- Section 4.3 demonstrates how MatSeek resolves the challenge of data integration described in Section 1.2.1
- Section 4.4 discusses how MatSeek differentiates from the previous research reviewed in Section 1.7.1, the limitations and future work plan.

4.1 System Architecture

Figure 4.1 illustrates the overall system architecture that comprises a set of key components on the server and the client sides respectively, and the search interface.

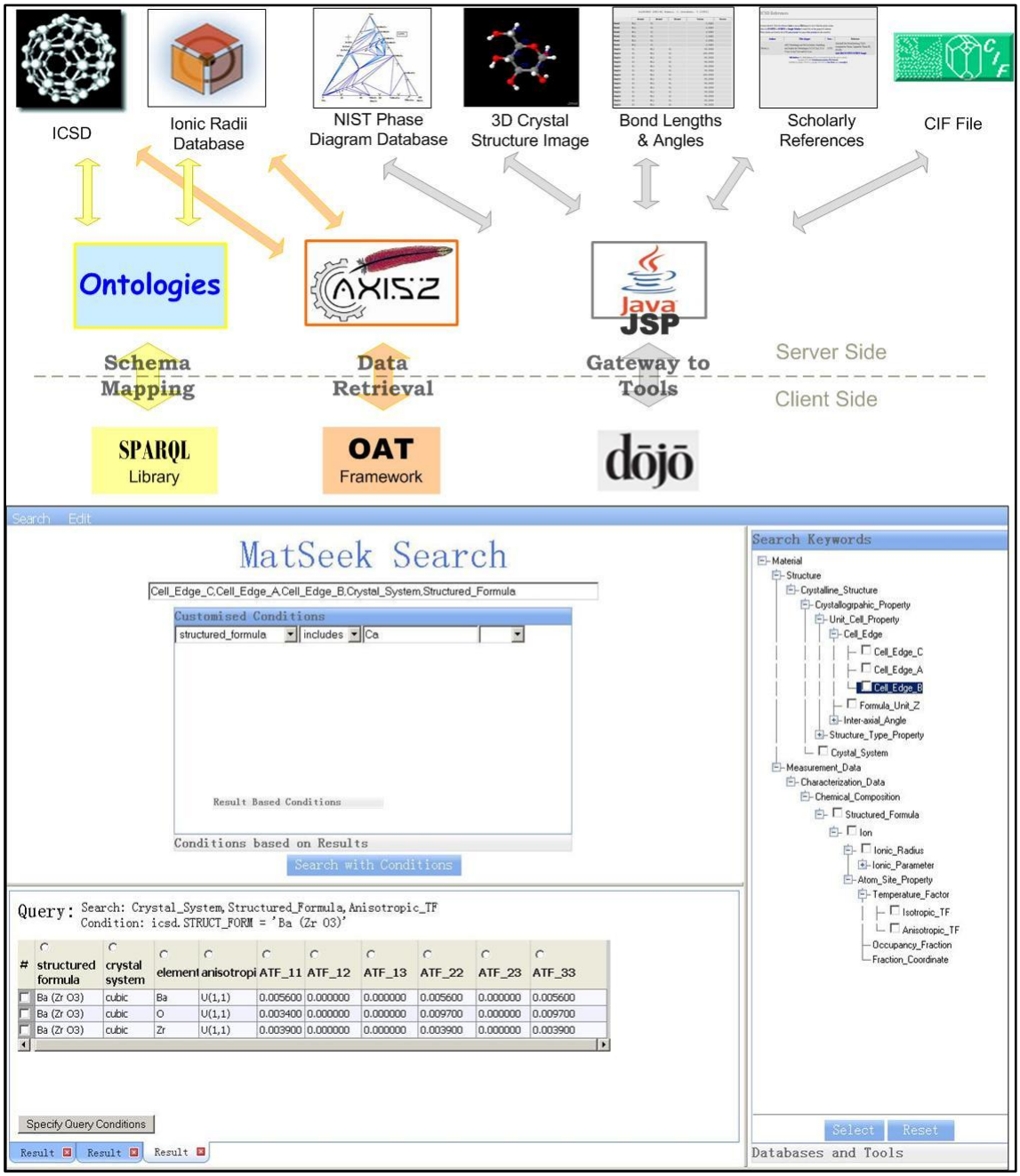


Figure 4.1: System Architecture

The key components on the **server side** are:

- The MatOnto ontology that is represented in OWL (Chapter 3)
- The Apache Axis2 [158] is the core engine for Web services. There is an independent and reusable web service as an entry point for querying databases including the Inorganic Crystal Structure Database (ICSD) and the Ionic Radii database
- A web application has been developed using JavaServer Pages [159] on top of the Apache Tomcat as a gateway to accessing the NIST Phase Equilibria Diagrams database, rendering 3D crystal structure images, calculating bond lengths and angles, locating and retrieving scholarly references, and exporting Crystallographic Interchange Files.

The key components on the **client side** on which MatSeek's search interface is based are:

- a SPARQL JavaScript library [160] supporting the querying of the ontologies on the server side
- the OpenLink AJAX Toolkit (OAT) Framework [161] enabling MatSeek to invoke the web service for data retrieval on the server side and present retrieved data on the client side. The OAT Framework is an open-source JavaScript-based library for browser-independent Rich Internet Application development. The Asynchronous JavaScript and XML (AJAX) programming [162] is a web development technique that enhances web pages' responsiveness, interactivity and usability.
- the MatSeek's user interface, which is rendered by the Dojo widget library. Dojo [163] is an open source JavaScript library, designed for the rapid development of AJAX-based applications and websites.

Figure 4.2 demonstrates MatSeek's search interface. On the RHS of the interface is an accordion widget that consists of two components, the *Search Keywords*, and the *Databases* and *Tools* panels. The former has a hierarchical structure of search keywords for browsing. The latter enables the users to access the phase diagram database and analysis tools and lists the available databases — the ICSD and Ionic Radii databases. Next, the *Search Panel* on top of the LHS consists of a textbox for targeted search keywords, an accordion widget for customising search conditions, and a button for invoking a search request. Finally, the *Result Panel* on the bottom of the LHS is in a tabbed-page format that displays each individual search result.

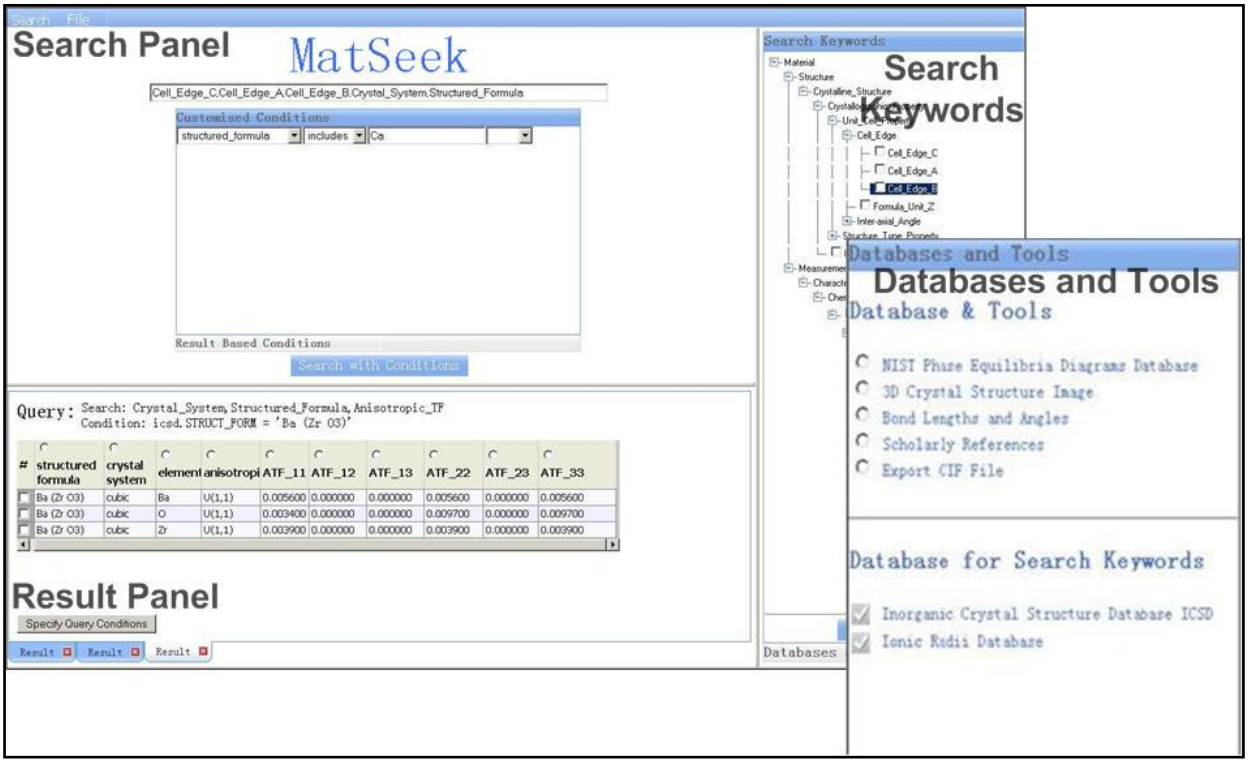


Figure 4.2 Search Interface

4.2 Technical Perspective

This section abstracts the technical details behind MatSeek in, (1) the manual mapping of the database schemas through a mapping ontology, (2) the dynamic construction of precise SQL query statements through an ontology modelling referential relationships and, (3) the systematic retrieval, correlation and integration of data retrieved from disparate but relevant databases through a correlating ontology. Thus, there are two ontologies required. One is MatOnto, the domain-specific ontology, for the schema mapping and data correlation and integration, while the other is the Referential Relationship Ontology developed and based on the structure of the relational database schema in the aspect of the referential relationship.

4.2.1 The MatOnto Ontology

The MatOnto described in Chapter 3 is an extensible ontology, based on the DOLCE [139] upper ontology, that aims to represent the structured knowledge about materials, their structure and properties and the processing steps involved in their composition and engineering. The MatOnto enables the mapping between the materials science database schemas. Figure 4.3 demonstrates a subset of the MatOnto — defining the aspects of the structural properties and measurement data. The blue and brown nodes are classes mapping onto the entity attributes of the ICSD and Ionic Radii database schemas, respectively, while the purple arrow labelled owl:equivalentClass indicates that

the *chemicalElement* and *ion* classes have the same instance populations. The methodology for mapping between database schemas is discussed in Section 4.2.3.

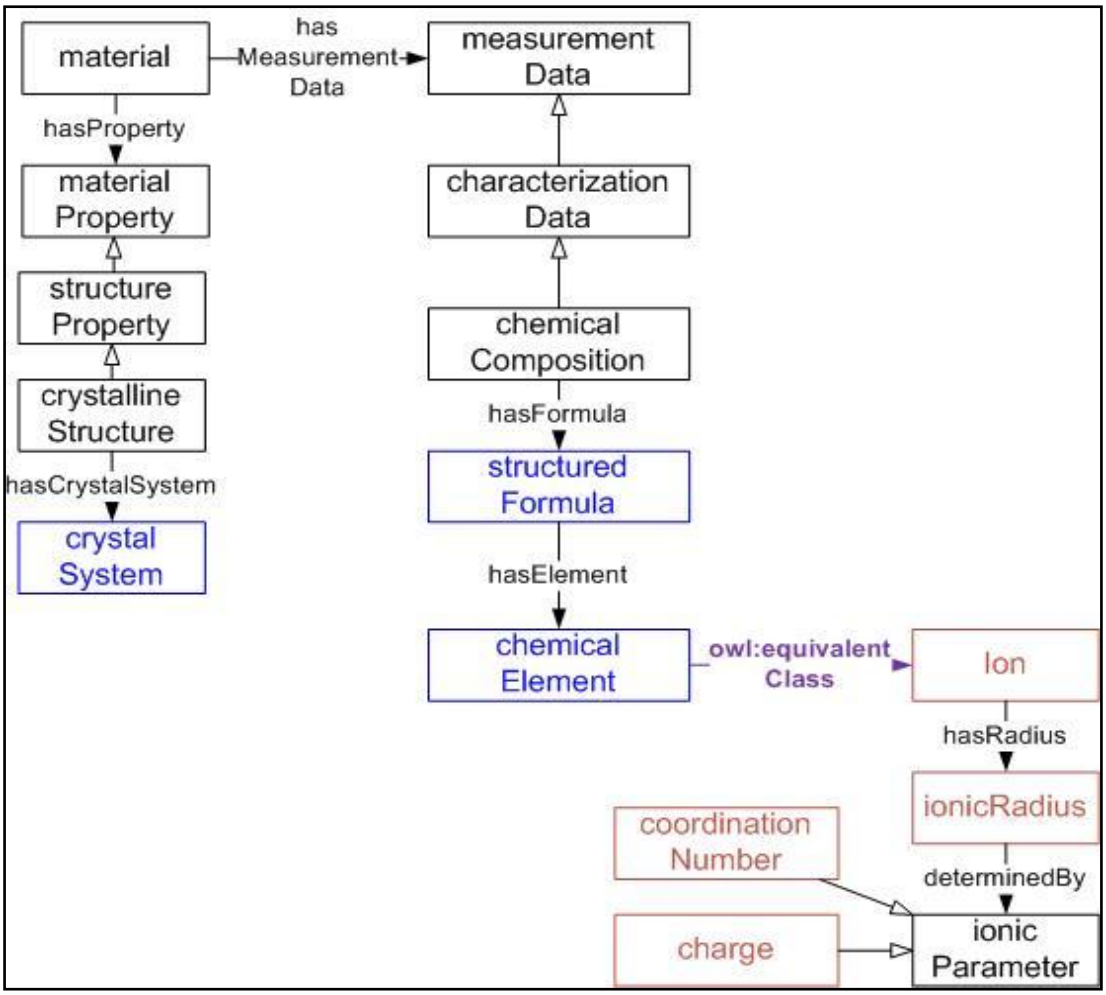


Figure 4.3 MatOnto's Components about Measurement Data and Ionic Radii

4.2.2 Referential Relationship Ontology

The Referential Relationship ontology models the referential relationships between the entities within a relational schema. Figure 4.4 demonstrates the ontological structure consisting of four interlinked classes — database, entity, keyAttribute and non-keyAttribute. In particular, the keyAttribute class links to itself in a bi-directional way, because it contains populations of primary and foreign keys. This ontology enables MatSeek to, (1) infer referential relationships between the entities through foreign keys from the entity attributes mapped onto search keywords, (2) construct a SQL query statement dynamically and accurately that includes the attributes and entities, and that also joins the entities and, (3) search with controlled keywords, thereby providing an intuitive, Google-like user-friendly search interface.

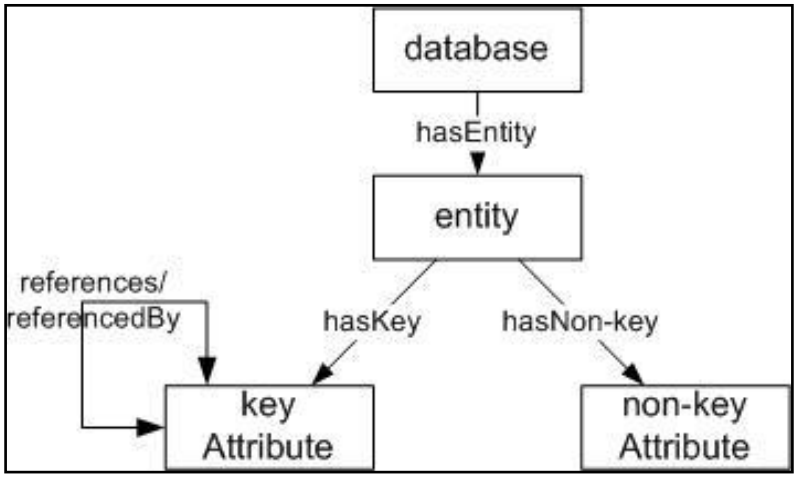


Figure 4.4: Referential Relationship Ontology

4.2.3 Manual Mapping of Database Schemas

Human effort is required to understand the meaning of the database schema terms and populate them under the MatOnto classes accordingly as the instances. Each of these instances consists of two strings divided by a dot. The former is the entity name in small letters, while the latter is the attribute name in capitals, which belongs to the entity, for example, p_record.EL_SYMBOL, where p_record is an entity and EL_SYMBOL is the attribute. Additionally, different database terms are mapped if they are within the same classes or in the different classes, if there is a semantically equivalent relationship between them.

Given the search keywords/ontological classes, through the populated ontology, the corresponding database terms can be identified automatically through a simple SPARQL query that identifies the instance of a specified class. The prefixes of those mapped items indicate the entities to which those items belong. Through the referential relationship ontology, the corresponding databases are also identified. As a result, an SQL query statement can be constructed dynamically.

4.2.4 Dynamic Construction of SQL Query Statements

As the involved databases, entities and attributes are identified, the next step is to determine which database should be queried first. In this case, the ICSD and Ionic Radii databases are about chemical compounds and ions, respectively. That means there is a composite relationship between a compound and its ions. A compound must be queried to identify the ions. As a result, query the ICSD database first, and then the Ionic Radii database with the elements of the resulting compounds.

Even though the involving databases, entities and attributes have been identified, these are still not sufficient to develop an accurate SQL query statement, because those entities may not be referenced directly within the schema. The unknown entities, as gaps between those known entities, must be

identified with the key attributes for developing the *join* conditions within a SQL query statement. Through the *keyAttribute* and *entities* classes of the referential relationship ontology, one of the known entities is used as the starting point to find its adjacent entity through the key attributes, and then both entities are grouped as an element and put into an array. Next, the newly found entity is the starting point to repeat the same process until all other known entities are reached. As a result, every two elements have an overlapped entity. The final result is an array that includes a sequence of entity-pair elements.

A hierarchical tree structure is developed using the array's elements through the overlapping entities between the array elements to find the routes between the known entities. Every route includes a chain of entities. Through the referential relationship ontology, the associated key attributes with their entities are identified for constructing the *join* conditions. Eventually, a SQL query statement can be constructed precisely.

4.2.5 Data Correlation and Integration

The process of correlating and integrating data retrieved from disparate but relevant databases is automatic and achieved through the MatOnto, the correlating ontology. First, after querying the ICSD database, the returned data set comprises rows of data items with a row of corresponding attributes names. Second, the attribute names are mapped onto the names of the ontological classes, as in Section 4.2.3. The data items are then populated into the classes as instances with auxiliary instances to bind those data items belonging to the same row together. Third, when the system identifies there is an equivalent relationship between the chemical element and ion classes through owl:equivalentClass, it replicates the instances of the former to the latter. Fourth, the system constructs a new query statement for the other database — the Ion Radii database, as in Section 4.2.4. The search attributes of the new statement include the *ion* instances. Fifth, after querying the Ionic Radii database, the system maps the returned attributes names and populates the associated data items into MatOnto accordingly, also with the auxiliary instances. Finally, the system correlates and integrates the data items from both databases through the equivalent and auxiliary instances.

4.3 Implementation and User Interfaces

This section describes how MatSeek implements a search request across the ICSD and Ionic Radii databases simutaneously. It returns the aggregated search results as input to the NIST Equilibria Diagrams database and analysis tools.

Consider once again, the materials scientist from the Case Study in Section 1.3. After logging onto MatSeek, the materials scientist wants to search for compounds that include tungsten and belong to the cubic crystal system. The compounds' crystal structure information is in the ICSD database,

while the ionic conductivity data is stored in the Ionic Radii Database. Figure 4.5 demonstrates the process, (1) a user selects search keywords from the *Search Keyword* on the RHS panel, (2) as the user confirms the selected keywords, MatSeek dsiplays the keywords in the textbox on the LHS panel. The keywords correspond to the entity attributes of the ICSD and Ionic Radii database schemas respectively and, (3) the user-customised search conditions are on the accordion container below the textbox.

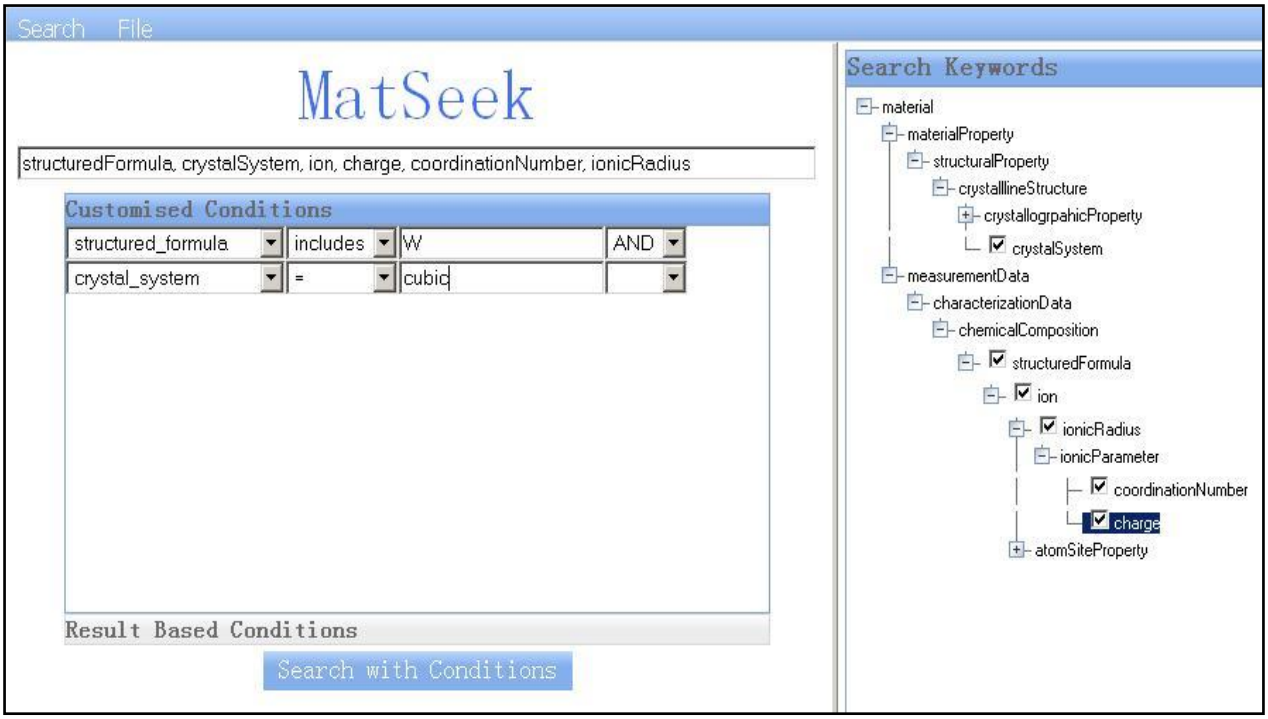


Figure 4.5: Search Request

When the user hits the *Search* button, MatSeek responds to the search request in the following way. First, it maps the search keywords onto the entity attributes in the database schemas through the MatOnto. Next, it works out the referential relationships between the involving entities through the Referential Relationship Ontology, and then constructs precise SQL statements dynamically. Finally, it queries both databases, correlates and integrates the returned data, and presents it to the user. The technical details are detailed in the following sections.

4.3.1 Database Schemas Mapping

Figure 4.6 demonstrates the steps for the mapping between the MatOnto's ontological classes and the entity attributes from the schemas of the ICSD and the Ionic Radii databases.

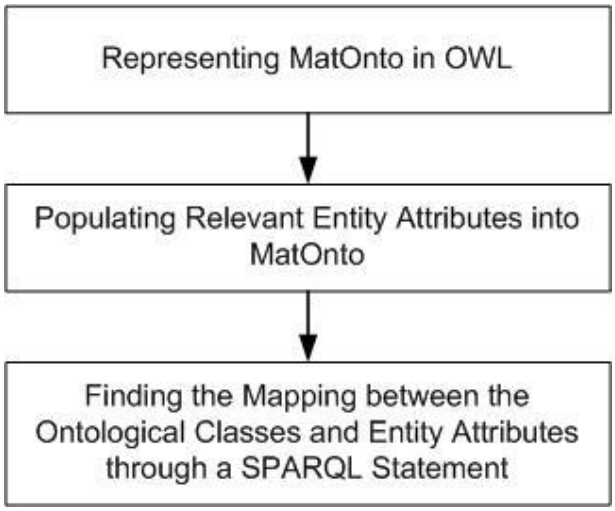


Figure 4.6: Steps for the Mapping between MatOnto and Database Schemas

1. MatOnto has been represented in the Web Ontology Language (OWL) as a data model through Protégé 4. Figure 4.7 demonstrates the imported ontologies, including DOLCE, EXPO, the JACS system, MPEG-7 ontology, Gruber and Olsen's Engineering Mathematics Ontology, the SWRC ontology, the SUMO ontology and W3C's Time Ontology described in Chapter 3, while Figure 4.8 demonstrates the entries of the classes and object properties.

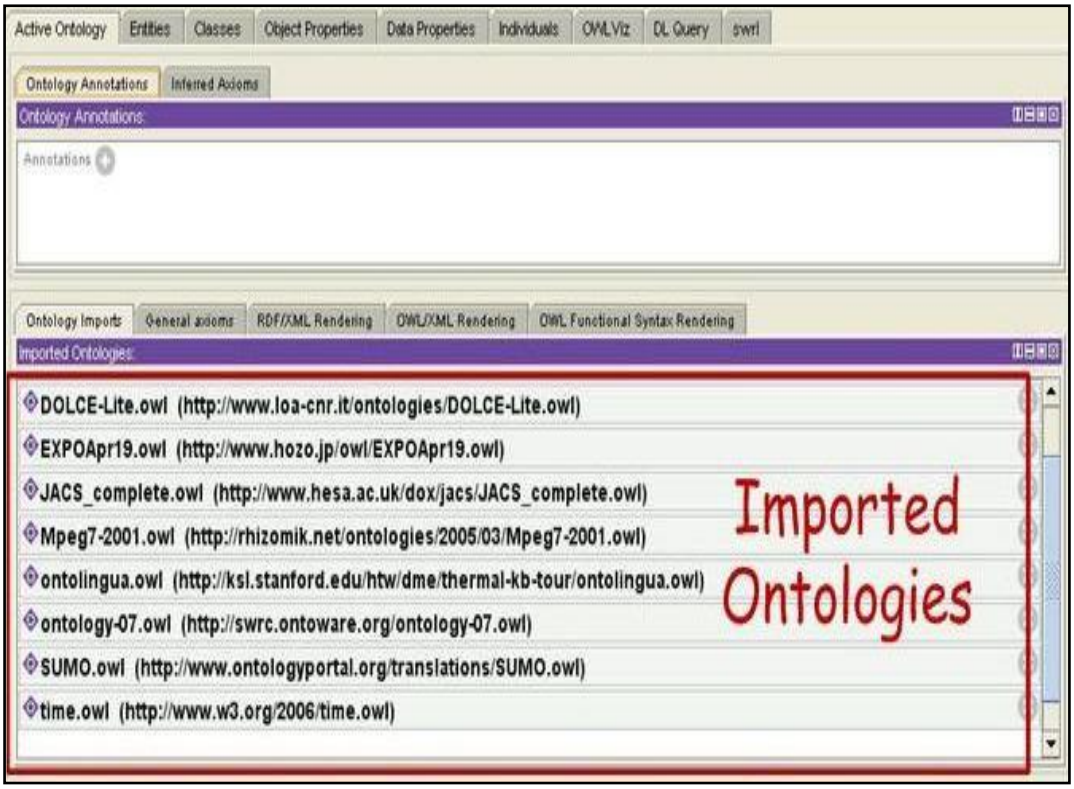


Figure 4.7: Imported Ontologies

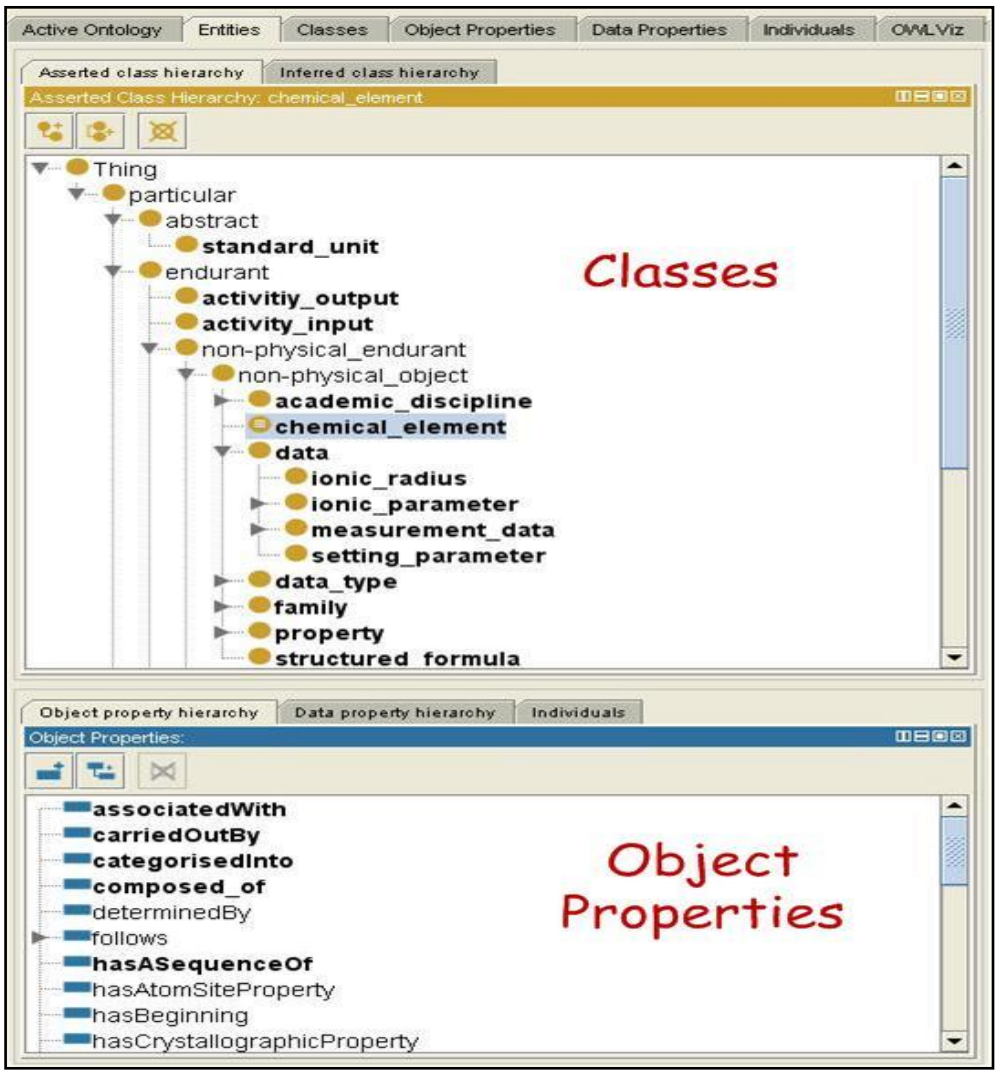


Figure 4.8: MatOnto's Classes and Object Properties

2. The names of the relevant entity attributes from both databases have been populated into the semantically corresponding ontological classes, also as the instances through Protégé 4. Figure 4.9 conceptualises the mapping. For example, the *structureFormula* class matches the ICSD attribute *icsd.STRUCT_FORM*, while the *ionicRadius* class matches the Ionic Radii database attribute *ionic_radii.IONIC_RADIUS*. Additionally, the predicate owl:equivalentClass bridges the semantic gap between both databases through the equivalent relationship between the *chemicalElement* and the *ion* classes in terms of symbolic representations. Figure 4.10 demonstrates the entry of the *ionic_radii.ION* instance under the class *ion*. It indicates the implemented equivalent relationship between the ion and *chemical_element* classes highlighted on the Classes panel, while it indicates the *ionic_radii.ION* is the same as the *p_record.EL_SYMBOL* on the LHS of the *Instance Details* panel. Furthermore, it also indicates the ionic_radii.ION relates to *ionic_radii.IONIC_RADIUS* on the RHS of the same panel.

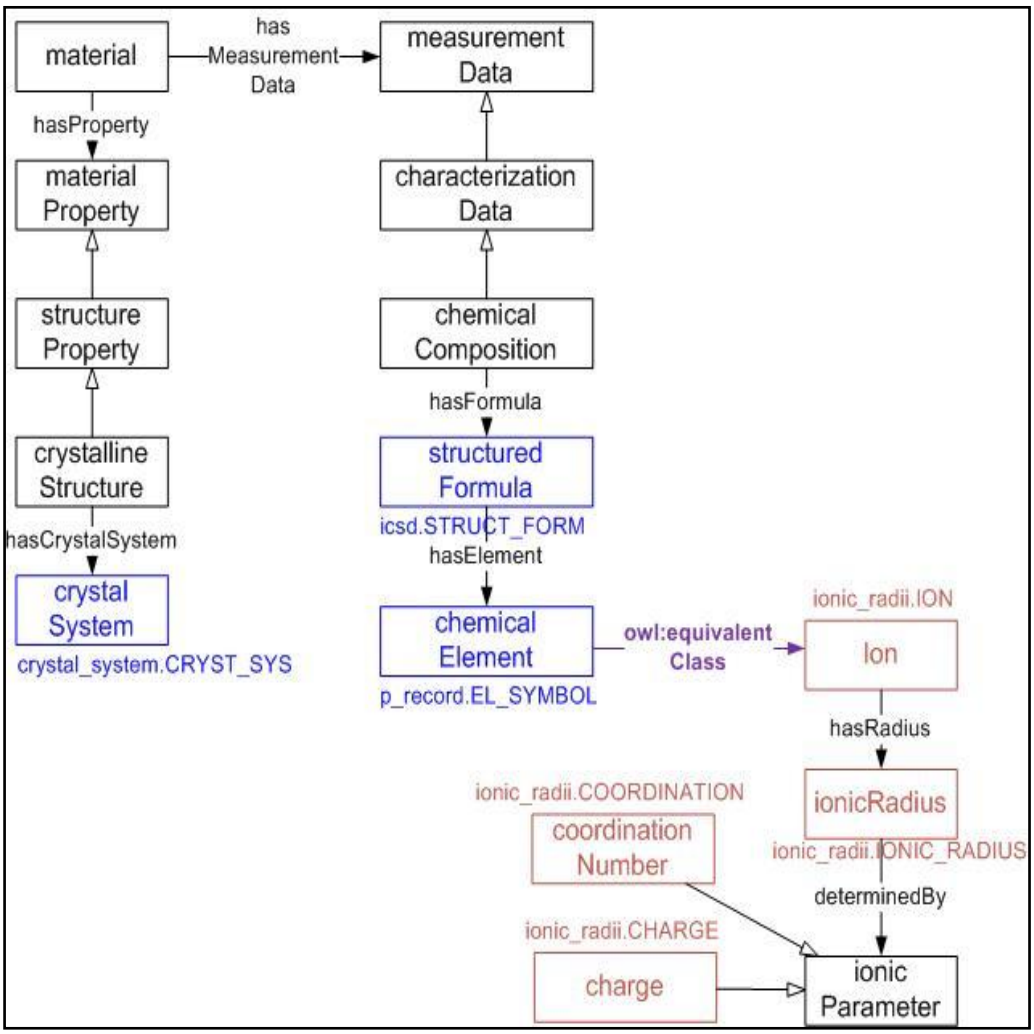


Figure 4.9: Database Metadata Mapping through MatOnto

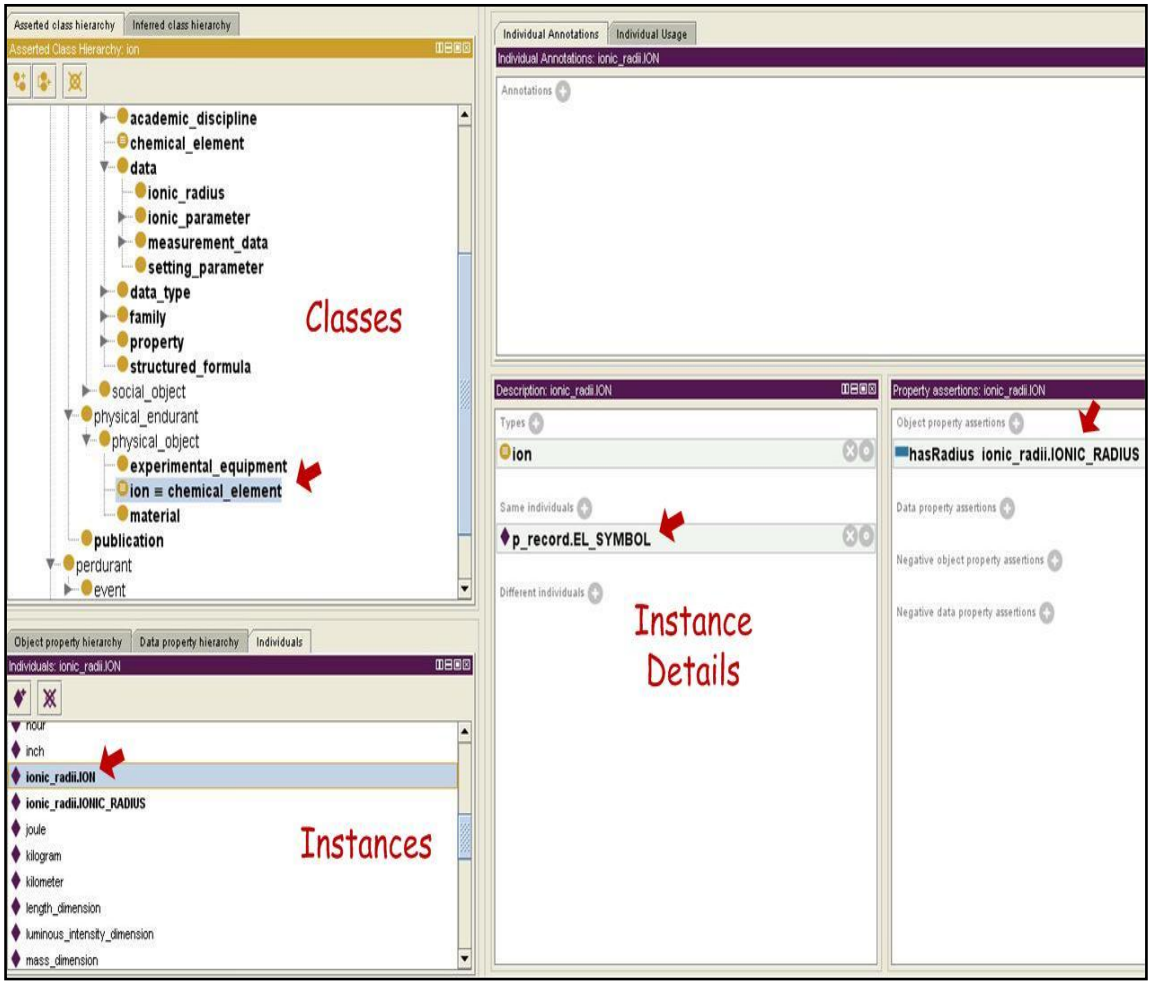


Figure 4.10: MatOnto's Instances Details

3. The mapping between the search keywords (ontological classes) shown in Figure 4.5 and the database terms (the instance names) is through the execution of the following SPARQL statement on the populated MatOnto by the SPARQL JavaScript library [160]. The *search_keyword* term at the *WHERE* clause of the SPARQL statement is a variable that is assigned the search keyword. For example, as search keyword *structuredFormula* is assigned to the variable, the returned *entittyAttribute* is *icsd.STRUCT_FORM* following executing the SPARQL statement.

   PREFIX MatOnto: <http://localhost:8080/Onto/MatOnto/MaterialsOntology.owl#>

   PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

   SELECT ?entityAttribute

   WHERE { ? entityAttribute rdf:type search_keyword}

4.3.2 Finding Referential Relationships

Figure 4.11 demonstrates the steps of finding the referential relationships between targeted entities.

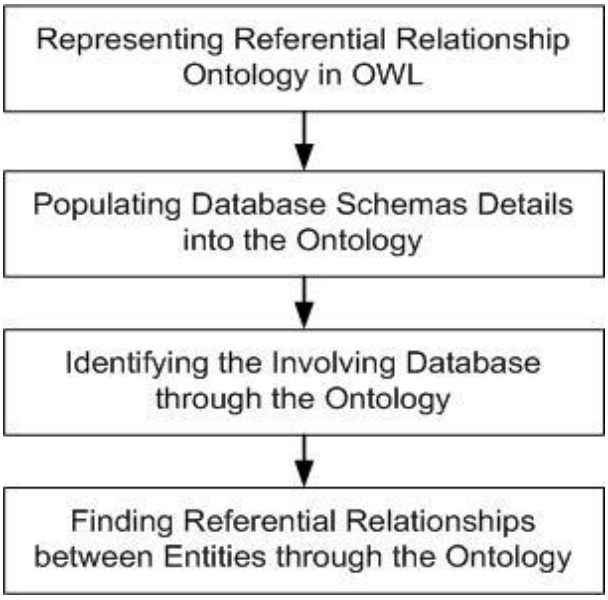


Figure 4.11: Steps for Finding Referential Relationships between Entities

1. The Referential Relationship Ontology has been represented in OWL as a data model through Protege 4. Figure 4.12 demonstrates the entries of classes and object properties.

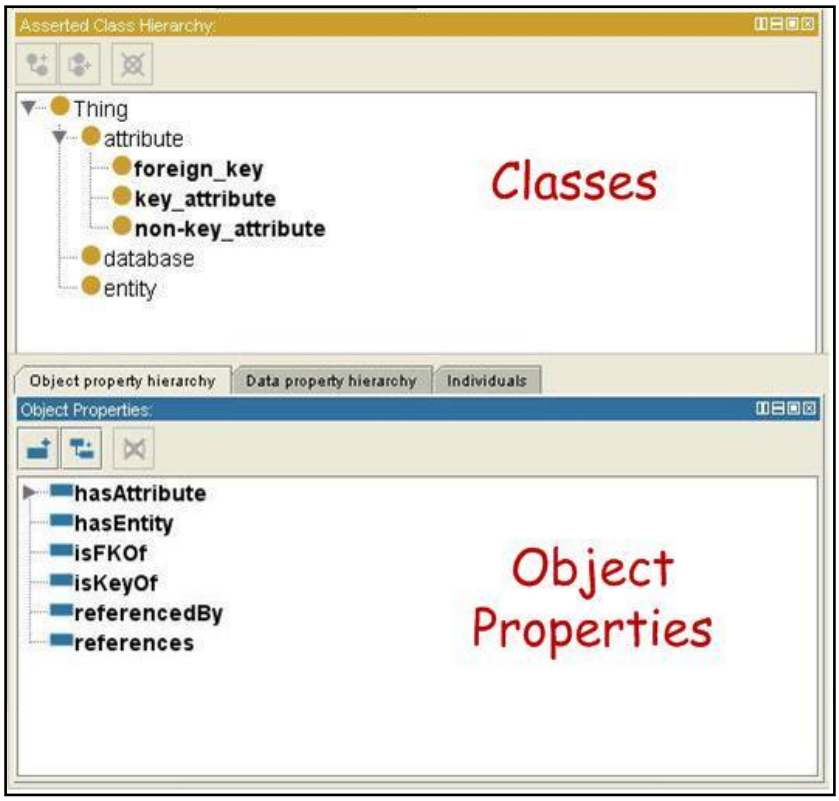


Figure 4.12: Classes and Object Properties of the Referential Relationship Ontology

2. The names of databases, entities, key and non-key attributes have been populated into the ontology accordingly. Figure 4.13 demonstrates the details of the *DB_ionic_radii* instance.

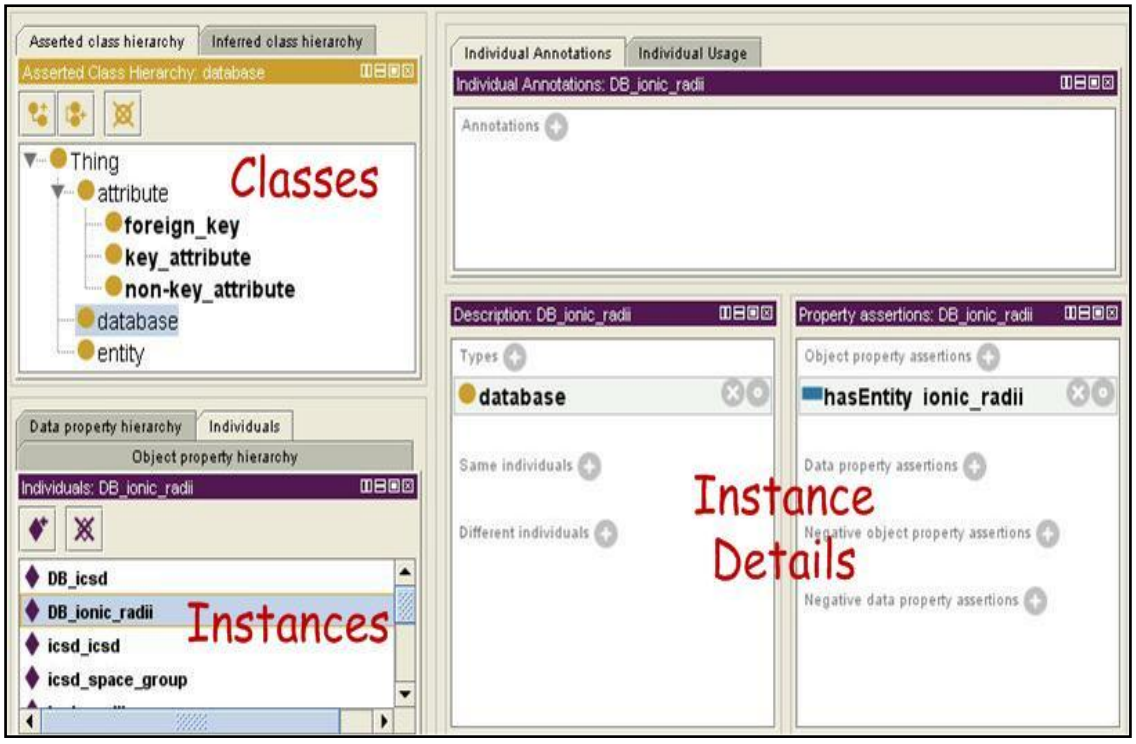


Figure 4.13: Instance Details

3. Because the mapped entities attributes have been identified in Section 4.2.3, MatSeek identifies the databases to which those attributes belong by executing the following SPARQL statement on the Referential Relationship Ontology through the SPARQL JavaScript library. The *prefix_name* variable in that statement is assigned the prefixes of the mapped attributes to identify the database. Next, MatSeek constructs a SQL query statement for the ICSD database (because there exists an implict whole/part relationship between *icsd.STRUCT_FORM* and *ionic_radii.ION*). MatSeek has to work out the referential relationships between the involving entities in order to construct an accurate query statement.

```
PREFIX RROnto:< http://localhost:8080/Onto/RROnto/RROnto.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT  ? db
WHERE {  ? db                  rdf:type              RROnto:database .
         RROnto:database       RROnto:hasEntity      RROnto:entity .
         prefix_name           rdf:type              RROnto:entity}
```

4. Through the ontological predicates: *references* and *referencedBy* between the *key_attribuite* instances shown in Figure 4.4, MatSeek identifies the referential relationships between the

entities and then constructs an accurate SQL query statement dynamically. The pseudo-algorithm is as follows with the assumption that all the entites are interconnected within a database.

a. Given the mapped entity attributes within the ICSD database: *icsd.STRUCT_FORM*, *crystal_system.CRYST_SYS*, *p_record.EL_SYMBOL*, MatSeek
   i. identifies the *icsd*, *crystal_system* and *p_record* entities are involved in this query through the attribute prefixes
   ii. queries the Referential Relationship Ontology with the starting entity (*icsd*) and a specified database (the ICSD database) with SPARQL statements through foreign keys in order to discover adjacent entities
   iii. groups the starting and adjacent entities as an element and puts it into an array —the table-pair chain array.

b. this process (steps ii and iii) will repeat with the resulting adjacent entity as the starting point until all the target entities are all reached. The resulting array includes the following elements: (*icsd, p_record*), (*icsd, space_group*), (*space_group, space_group_number*), (*space_group_number, crystal_system*), (*icsd, mineral_name*), (*mineral_group*), (*icsd, structure_type*), (*structure_type, structure_type_statistics*), (*structure_type, wyckoff*) (*icsd, reference*), (*icsd, author_icsd*), (*author_icsd, author_table*).

c. because the table-pair array contains all the pair-entities, a tree with the starting entity – *icsd* – as a root is constructed using the pair-entity elements within the array. Figure 4.14 demonstrates the developed tree.

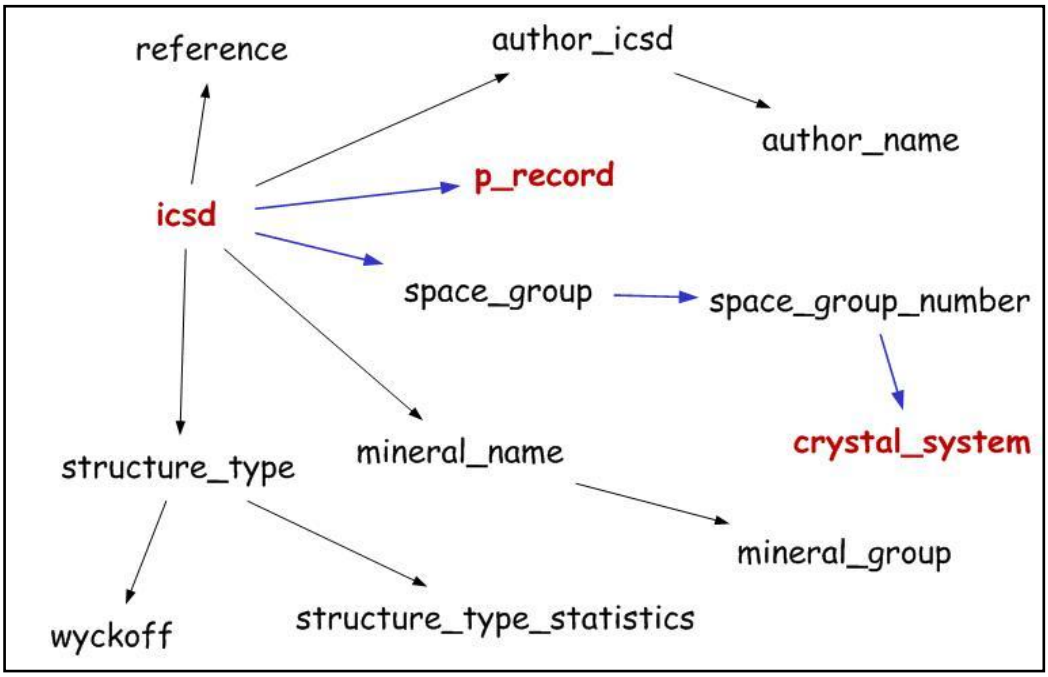


Figure 4.14: The Table-pair Tree

d. The tree is traversed to find routes from the starting entities/root to the target entities/leave nodes. Figure 4.14 also demonstrates the routes in blue between the starting entity, that is, *icsd*, and the targets —*p_record* and *crystal_system*.

e. Every route includes a chain of tables. Querying the Referential Relationship Ontology with a pair of interconnected tables to find out the foreign key(s) and corresponding primary key(s). Figure 4.15 demonstrates the chains of the entities with the associated key attributes.

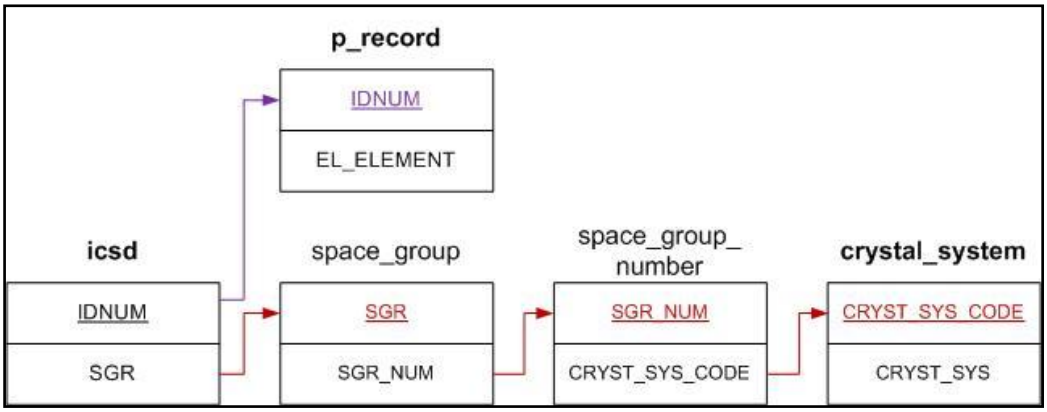


Figure 4.15: Referential Relationship Chains with Keys

f. As described in Section 4.2.1, there is an equivalent relationship between the *chemicalElement* and *ion* classes. Both classes have the *p_record.EL_SYMBOL* and *ionic_radii.ion* attributes from the ICSD database and Ionic Radii database as their instances, repectively. As a result, the SQL statements for querying the ICSD and Ionic Radii databases must include both attributes to correlate return data from both databases, even though those attributes are not required by a search request. Finally, MatSeek constructs the following query statement according to the search request shown on Figure 4.15.

SELECT icsd.STRUCT_FORM, crystal_system.CRYST_SYS, p_record.EL_SYMBOL

FROM icsd, p_record, crystal_system, space_group, space_group_number

WHERE

icsd.STRUCT_FORM LIKE '%W%' AND

crystal_system.CRYST_SYS='cubic' AND

crystal_system.CRYST_SYS_CODE=space_group_number.CRYST_SYS_CODE AND

space_group_number.SGR_NUM = space_group.SGR_NUM AND

space_group.SGR = icsd.SGR AND icsd.IDNUM = p_record.IDNUM

4.3.3 Data Retrieval, Correlation and Integration

Figure 4.16 demonstrates the steps for data retrieval, correlation and integration.

1. MatSeek invokes the database access web service hosted in Apache Axis2 on the server side using the web service JavaScript library of OAT Framework to query the ICSD database with the query developed in Section 4.3.2. The returned result set comprises rows of data items with a row of corresponding attribute names.
2. It maps those attribute names onto the ontological classes as in Section 4.3.1 using the query package of the JENA semantic web framework [164] on the service side, and populates the data items to those classes accordingly, using JENA's model package with the instances of an auxiliary class — ICSDResultRow — that binds the data items belonging to the same row together. Figure 4.17 demonstrates the population of the ICSD resulting data items and their binding through the ICSDResultRow instance — ICSDResultRow_1 and the predicate — hasDataItem (dot lines in blue).

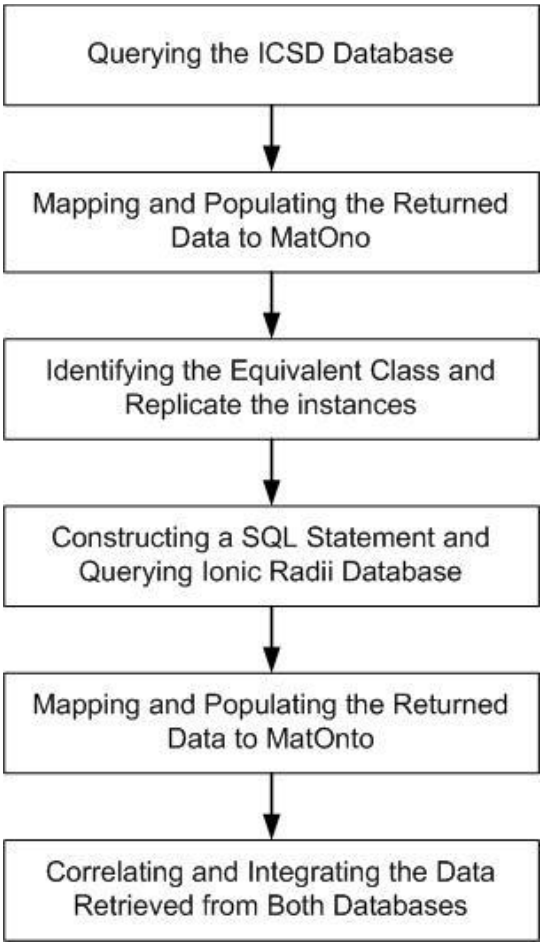


Figure 4.16: Steps for Data Retrieval, Correlation and Integration

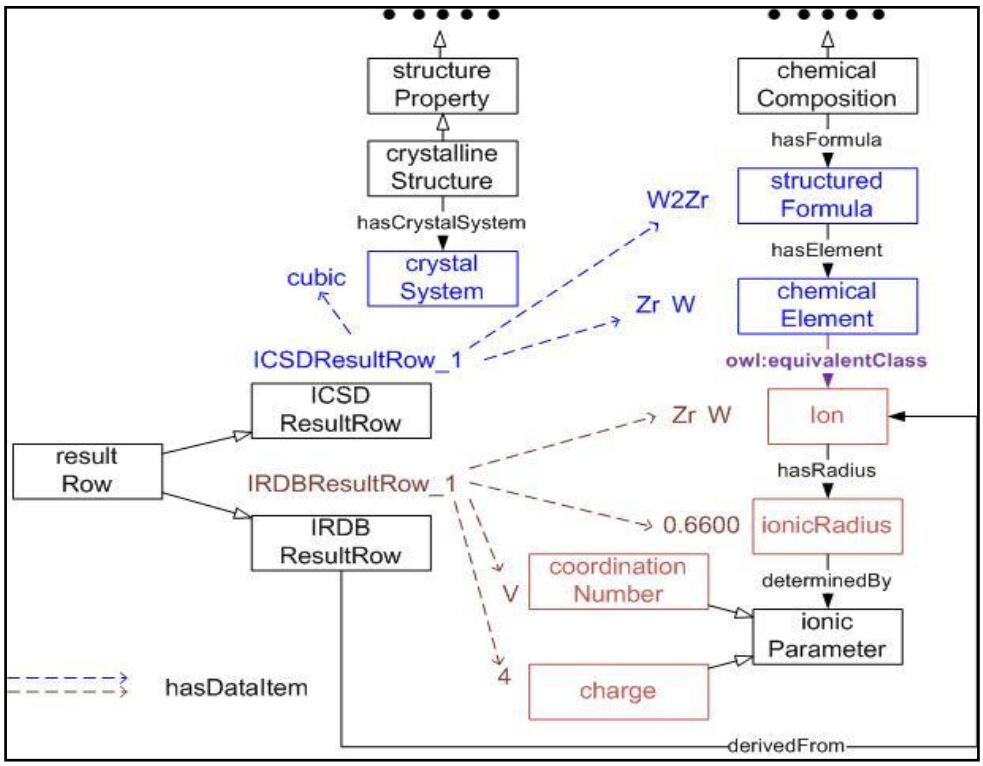


Figure 4.17: Population of Data Items from the Databases

3. It identifies the equivalent class of the *chemicalElement* class through the predicate *owl:equivalentClass* [165] with the following SPARQL statement, and replicates the instances in the class to the identified class —*ion* using JENA's query and model packages, respectively on the server side.

   PREFIX MatOnto: <http://localhost:8080/Onto/MatOnto/MaterialsOntology.owl#>
   PREFIX owl: <http://www.w3.org/2002/07/owl#>

   SELECT ?equivalentClass
   WHERE { MatOnto:chemicalElement owl:equivalentClass ?equivalentClass }

4. It constructs a SQL query statement that includes instances of class *ion* and the entity attributes mapped onto the search keywords about the Ionic Radii database, as in Section 4.3.1, but on the server side.

   SELECT ionic_radii.ION, ionic_radii.CHARGE, ionic_radii.COORDINATION, ionic_radii.IONIC_RADIUS
   FROM ionic_radii
   WHERE ionic_radii.ION = 'W' OR ionic_radii.ION = 'ZR'

5. It queries the Ionic Radii database, and gets through the mapping and populating processes same as Step 2, but with a different auxiliary class —IRDResultRow, as in Step 2. Figure

4.17 also demonstrates the population of resulting data items from the Ionic Radii database and their binding through the IRDBResultRow instance — IRDBResultRow_1

6. It correlates, integrates and retrieves the whole data collection excluding all the instances of those auxiliary classes from MatOnto with the following SPARQL statement using JENA's query package. Figure 4.18 demonstrates the returned collective resulting data in the *Grid* rendered by OAT Framework [161].

```
PREFIX matOnto:  <http://localhost:8080/Onto/MatOnto/MaterialsOntology.owl#>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:  <http://www.w3.org/2002/07/owl#>
SELECT      ?icsdResultRow          ?icsdDataItem
        ?icsdMetadataTerm
        ?struct_form        ?structFormMetadata        ?element
        ?irdbResultRow      ?irdbDataItem              ?irdbMetadataTerm
WHERE {
        ?icsdResultRow    matOnto:hasDataItem          ?icsdDataItem .
        ?icsdDataItem          rdf:type                      ?icsdMetadataTerm.
        ?icsdResultRow    matOnto:hasStructuredFormula  ?struct_form .
        ?struct_form      rdf:type                      structFormMetadata .
        ?struct_form      matOnto:hasElement            ?element .
        ?irdbResultRow    matOnto:derivedFrom           ?element .
        ?irdbResultRow    matOnto:hasDataItem           ?irdbDataItem .
        ?irdbDataItem          rdf:type                      ?irdbMetadataTerm  }
```

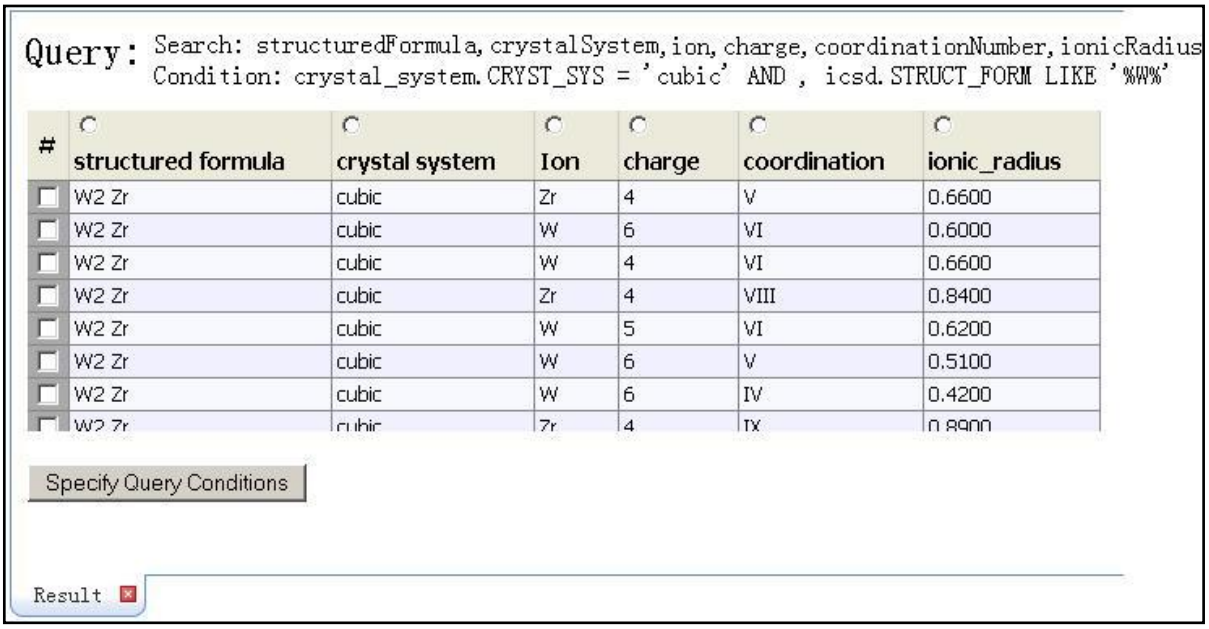


Figure 4.18: Search Results

In addition to the search functionality described above, MatSeek enables users to access the NIST Phase Equilibria Diagrams database (PED) and analysis tools through the accordion widget shown on the RHS of Figure 4.2. For example, users want to submit a particular compound retrieved by searching the ICSD and Ionic Radii databases to the NIST PED to see its stability at different temperatures and its 3D crystal structure. Figure 4.19 demonstrates the result page from the NIST PED, while Figure 4.20 demonstrates a rendered 3D crystal structure image. Figure 4.21 demonstrates the calculated bond lengths and angles, while Figure 4.22 demonstrates retrieved scholarly references.

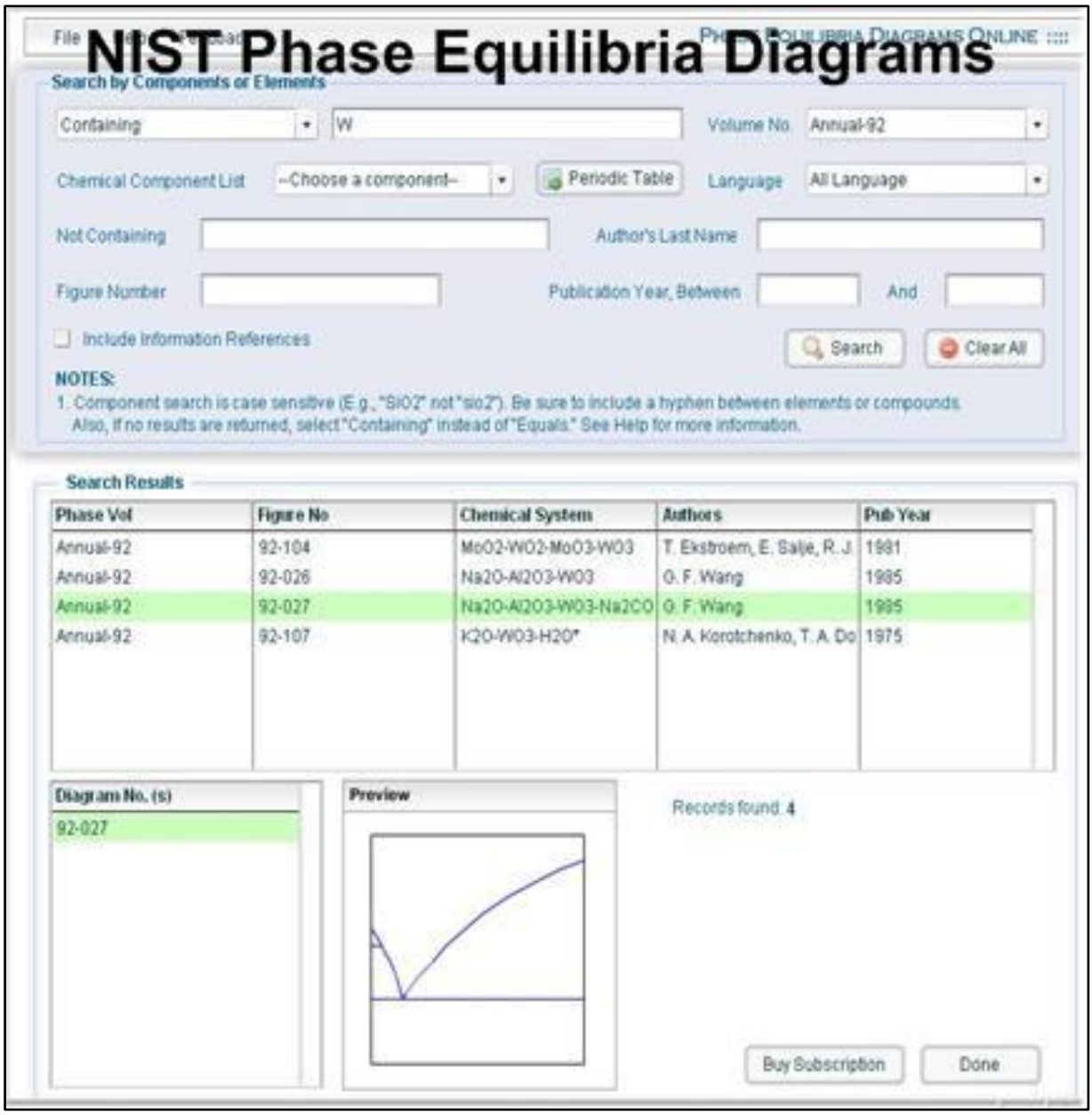


Figure 4.19: Search Result from the NIST database

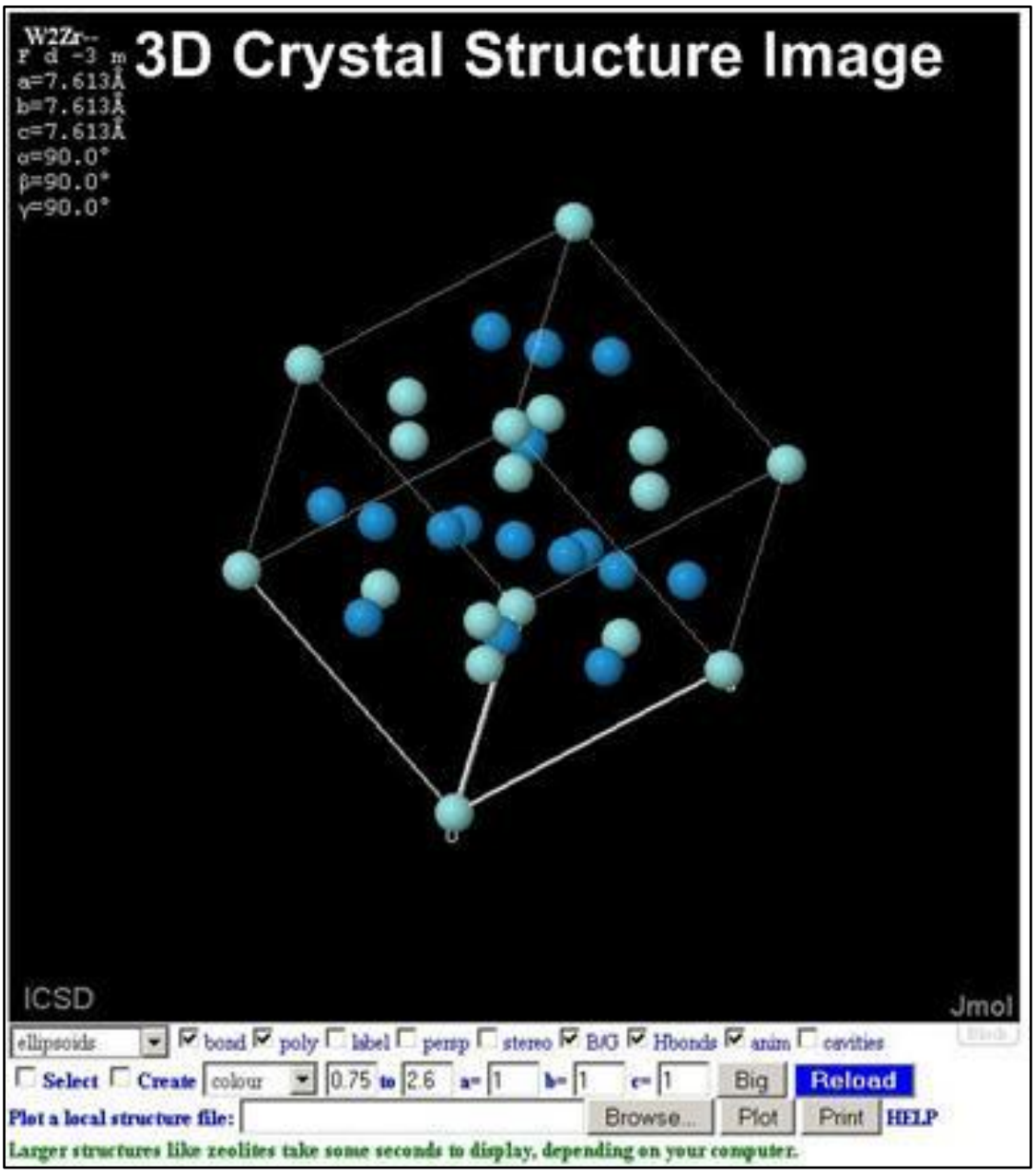
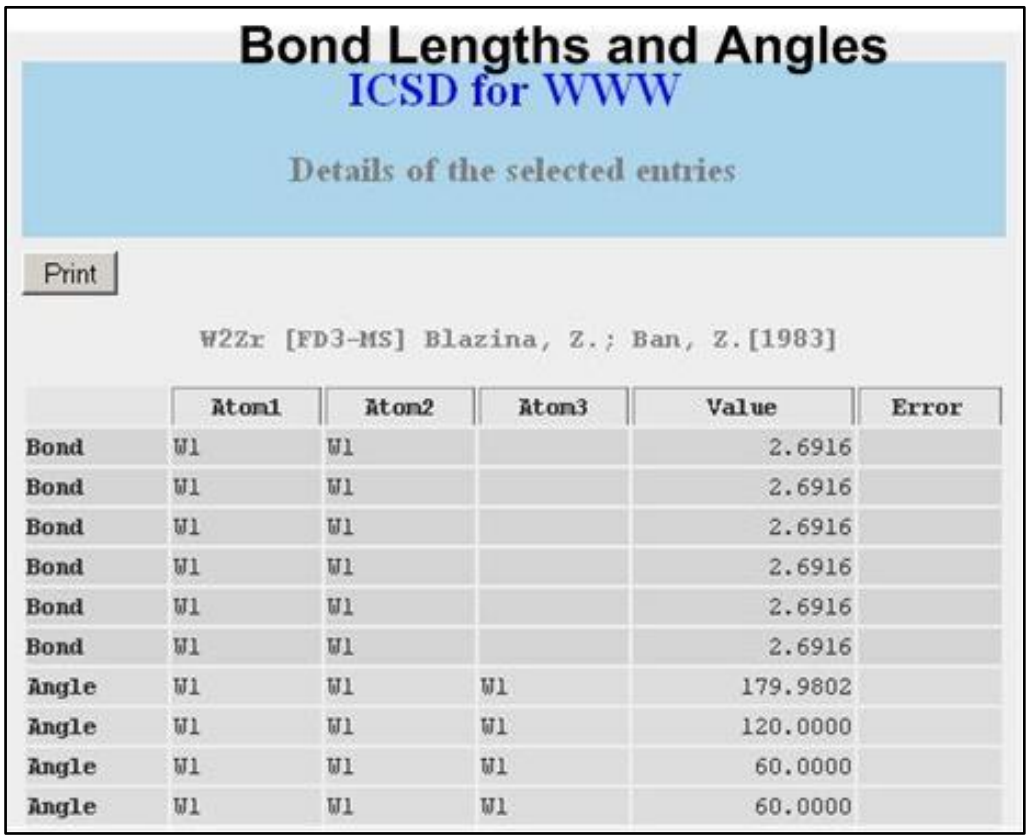


Figure 4.20: A Rendered 3D Crystal Structure Image



Bond Lengths and Angles

ICSD for WWW

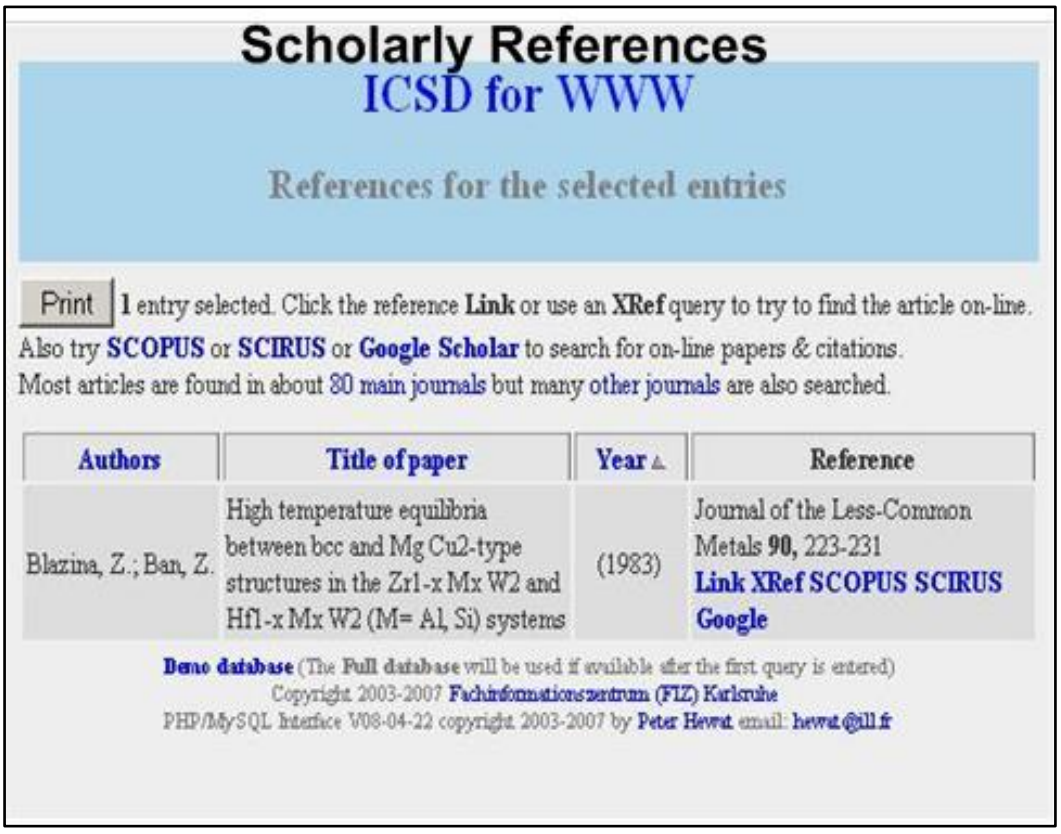
Details of the selected entries

Print

W2Zr [FD3-MS] Blazina, Z.; Ban, Z.[1983]

	Atom1	Atom2	Atom3	Value	Error
Bond	W1	W1		2.6916	
Bond	W1	W1		2.6916	
Bond	W1	W1		2.6916	
Bond	W1	W1		2.6916	
Bond	W1	W1		2.6916	
Bond	W1	W1		2.6916	
Angle	W1	W1	W1	179.9802	
Angle	W1	W1	W1	120.0000	
Angle	W1	W1	W1	60.0000	
Angle	W1	W1	W1	60.0000	

Figure 4.21: Calculated Bond Lengths and Angles



Scholarly References

ICSD for WWW

References for the selected entries

Print 1 entry selected. Click the reference **Link** or use an **XRef** query to try to find the article on-line.
Also try **SCOPUS** or **SCIRUS** or **Google Scholar** to search for on-line papers & citations.
Most articles are found in about 80 main journals but many other journals are also searched.

Authors	Title of paper	Year	Reference
Blazina, Z.; Ban, Z.	High temperature equilibria between bcc and Mg Cu2-type structures in the Zr1-x Mx W2 and Hf1-x Mx W2 (M= Al, Si) systems	(1983)	Journal of the Less-Common Metals **90**, 223-231 **Link XRef SCOPUS SCIRUS Google**

Demo database (The **Full database** will be used if available after the first query is entered)
Copyright 2003-2007 **Fachinformationszentrum (FIZ) Karlsruhe**
PHP/MySQL Interface V08-04-22 copyright 2003-2007 by **Peter Hewat** email: **hewat@ill.fr**

Figure 4.22: Retrieved Scholarly References

4.4 Discussion

4.4.1 User Feedback

We have evaluated MatSeek by deploying it within a team of materials scientists working within the AIBN at the University of Queensland. Feedback to date has been very positive. A survey of users conducted after a usability testing completed indicated that they were very impressed with the convenience of MatSeek, because of the easy access to integrated databases and the common MatOnto ontology. They requested the incorporation of further analytical and modelling tools within MatSeek's accordion widget. Further collaboration with the materials scientists is required to identify and embed the specific tools for statistical analysis, trend analysis, data mining, modelling, simulation and visualization. One of the major limitations identified by users is the lack of data in the publicly-available databases that we have incorporated. Commercial databases are more complete and comprehensive but outside the scope and budget of this project. Hopefully, over time, the culture of sharing materials science data through open access archives will become more widely adopted in the materials science community, thus this situation will improve.

4.4.2 Strengths

As far as we are aware, there are no other open source workbenches for materials science that are built on a combination of Semantic Web and Web 2.0 technologies. These technologies provide MatSeek with a relatively simple solution for database schema mapping and search interface and enhance the human-computer interactions. As discussed in Section 4.3.1, the mapping between the ontological terms and the entity attributes from the database schemas is implemented via a simple SPARQL statement rather than an inferencing engine (as Zhang has used in Section 1.7.1). The Referential Relationship Ontology enables MatSeek to search with controlled keywords, thereby resulting in an intuitive, Google-like and user-friendly search interface, as discussed in Section 4.3. Finally, Web 2.0 enables MatSeek to keep all of resulting data from the search requests in a series of tabbed pages within a single web-page, so they can easily be reused in subsequent queries.

4.4.3 Limitations and Future Work

MatSeek developed to-date is a working prototype that demonstrates the benefits of a single entry point to the key materials databases and analysis tools. However, further effort is required to improve the system's usability and robustness and to overcome existing limitations that include:

- The ICSD and NIST Phase Equilibria Diagrams databases used are for demonstration purposes, thus they have a limited number of data entries

- Human efforts are required to populate the names of the databases, entities and attributes as instances into the MatOnto ontology from the database schemas. Ideally, the uploading and mapping of new database schemas could be streamlined via a web interface
- The current available tools are mainly for the analysis of crystal structure data. There are many other analytical and modelling services that could usefully be incorporated
- MatSeek has a shallow integration with the NIST Phase Equilibria Diagrams database.

Future plans include:

- providing access to:
    - additional materials science databases in different formats including the XML and object-oriented databases, and RDF triple stores
    - additional analysis tools for statistical analysis such as the R [166] — an open source software environment for statistical computing and graphics, in addition to data mining tools such as WEKA [167] — the data mining with open source machine learning software
    - modelling and simulation tools and visualization tools
- Semi-automating the process of populating names of databases, entities and attributes from relational database schemas as instances into the MatOnto ontology.

4.5 Summary

In this chapter, we have described MatSeek as a federated search interface underpinned by MatOnto as discussed in Chapter 3, to key materials science databases and analytical tools. MatSeek is designed to enable materials scientists, through a single web-based platform, to search across databases containing crystal structure data, ionic conductivity data and phase stability data, render 3D crystal structure images, calculate bond lengths and angles, retrieve relevant scholarly references and export Crystallographic Information File. MatSeek appeals to the materials scientists in terms of intuitiveness and convenience. On one hand, the Google-like GUI makes the learning curve shallow. On the other hand, MatSeek enables, (1) disparate databases data to be correlated, integrated and presented collectively, (2) all of the resulting data returned by different search requests to be kept within the same page for better interactivity and, (3) the process from searching to analysing to be carried out within a single platform, thereby enhancing the scientists' productivity and capability in the precise process of assimilating materials data.

Chapter 5 SCOPE –Scientific Compound Object Publishing and Editing System

5.0 Introduction

The ability for sharing scientific data with the provenance information is critical to the verification and repeatability of scientific results, the peer review of scientific methodology and the progress of science. Recently, many organizations and funding bodies are actively encouraging or even mandating the publication of scientific data together with traditional scholarly publications across many domains. However, there are a number of barriers, as discussed in Section 1.2.2, that need to be overcome to get scientists to publish their datasets.

The WSBPEL workflow management system discussed in Section 2.5 enables systematic capture of data, metadata and provenance information generated by the scientific experiment workflow. Although, it is a very important initial step to assist scientists to publish their scientific results, there remain barriers to be overcome, a lack of simple tools for publishing data with provenance information; the concern with intellectual property rights and a lack of standards for publishing datasets with the provenance.

This chapter introduces the Scientific Compound Object Publishing and Editing system (SCOPE) that intends to remove those remaining barriers. The SCOPE system is designed to enable scientists to easily author, edit and publish scientific compound objects complying with the Open Archives Initiative Protocol – Object Exchange and Reuse (OAI-ORE). Scientific compound objects enable scientists to encapsulate the various datasets and resources generated or used during a scientific investigation process within a single compound object for publishing and exchange. OAI-ORE is an emerging standard to make the information within compound/complex digital objects discoverable, machine-readable, interoperable and reusable.

The SCOPE system plans to:

- provide scientists with multiple-grained levels of views of highly complex scientific workflows, so they can easily comprehend the science of the workflows and keep the confidential information from unauthorised viewers
- enable them to author a scientific compound object that:
    - incorporates the internal experimental data with the provenance information from the workflows and external digital objects with the metadata discoverable via the Web
    - is self-contained and explanatory with the IP protection
    - is assured to be widely disseminated on the Web.

The remainder of this chapter is structured as follows, Section 5.1 describes the system architecture, Section 5.2 describes the technical details. Section 5.3 presents a more specific scenario for evaluation and testing from the case study discussed in Section 1.3. Section 5.4 describes the implementation and user interface, Section 5.5 concludes with user feedback, limitations and future work plans.

5.1 System Architecture

Figure 5.1 illustrates the overall system architecture. It consists of:

- the Provenance (relational) database that contains the provenance information
- the Semantic Layer that contains D2R Map files and the D2R Processor [168], both of which convert the relational instances to the RDF triples
- the SWRL.OWL file contains the provenance information in RDF triples and the SWRL inference rules
- the provenance visualization tool —JGraph [169] and Jena [164] are used to convert an RDF graph into an image consisting of nodes (objects/classes) and arcs (relationships/properties)
- the Algernon rule-inference engine [170] —for inferring new indirect relationships based on the SWRL rules
- the SCOPE Authoring and Publishing Platform, which has four components:
    - the Provenance View
    - the Web Browser
    - the Publishing Interface
    - the metadata input and editing interface.

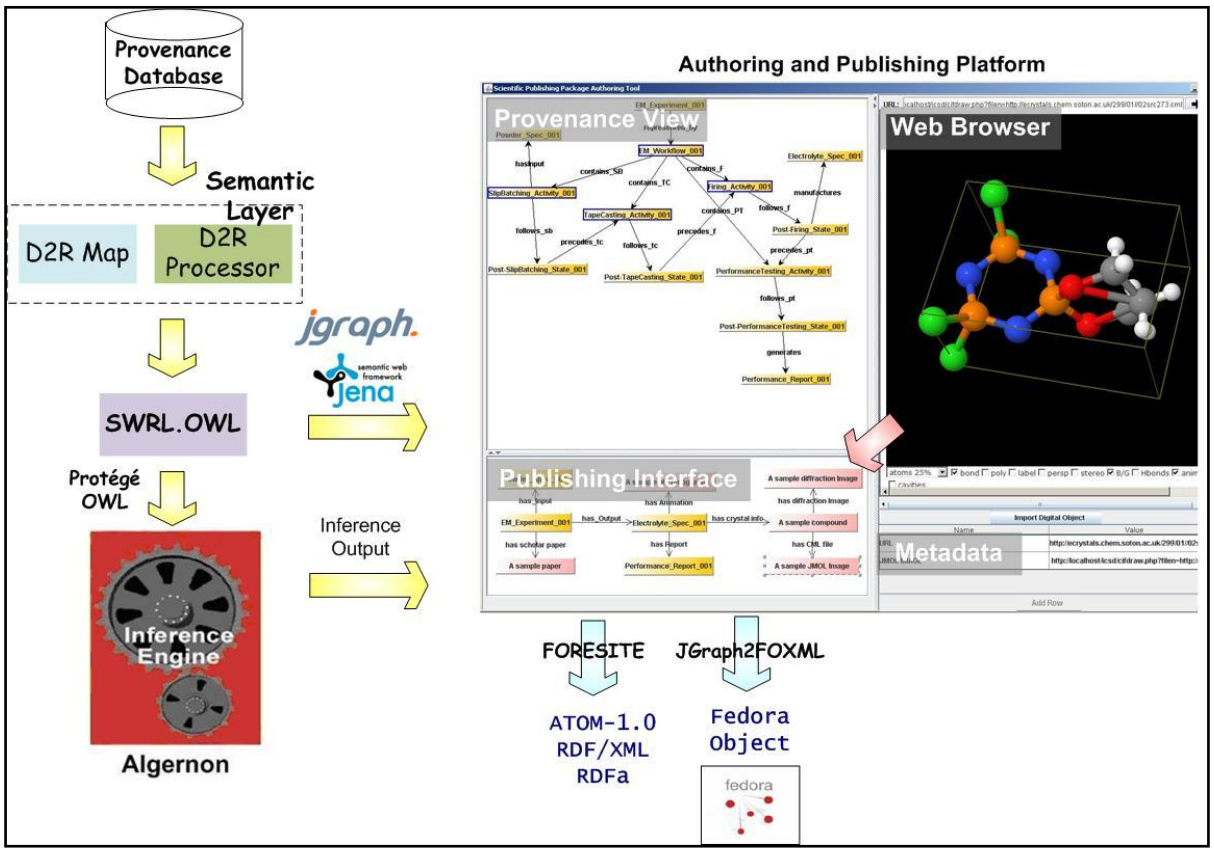


Figure 5.1: System Architecture

5.2 Technical Details

This section describes the technical details of the SCOPE system.

5.2.1 Access Control

Access controls are imposed on the Provenance view of the Authoring and Publishing Platform on the RHS of Figure 5.1. The granularity of the view depends on user privileges and access policies, enforced and defined by Shibboleth[171] and XACML [172].

To enforce the inter-institutional authentication and access control, Shibboleth, a centralized identity and authorization mechanism developed by the NSF Middleware Initiative, was adopted and incorporated within this system. Shibboleth is standards-based, open source middleware software that provides Web Single SignOn (SSO) across or within organizational boundaries. Figure 5.2 demonstrates the two primary components of Shibboleth, the Identity Provider (IdP) and the Service Provider (SP). The IdP maintains user credentials and attributes. Upon request, the IdP will assert authentication and attribute statements to requesting parties, specifically the SPs. The SP then uses predefined-XACML policies to control access to this system and the fine-grained provenance views on the Provenance View (described in Section 5.2.2.1.3)

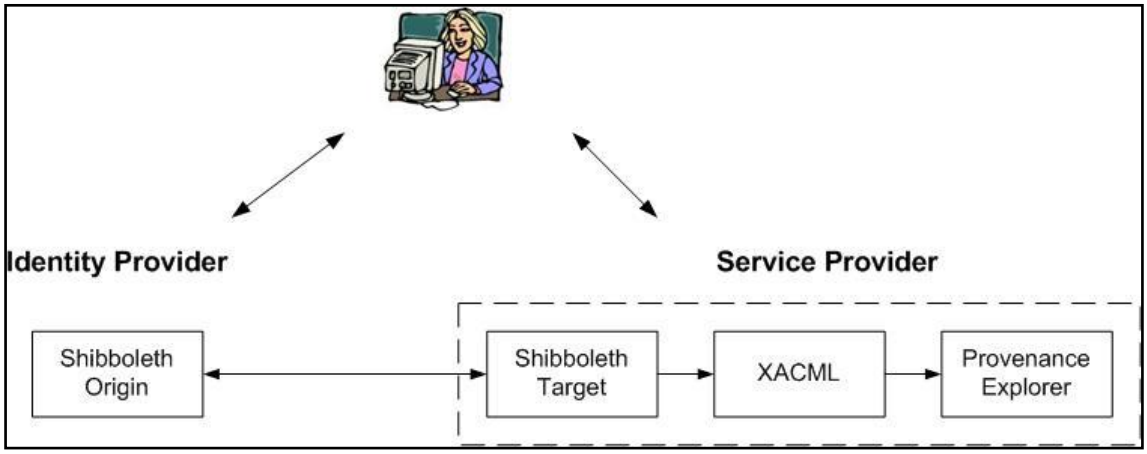


Figure 5.2: Authentication and Authorization System Architecture

5.2.2 The Authoring and Publishing Platform

The Authoring and Publishing Platform, at the top of the RHS of Figure 5.1, is the graphical user interface GUI powered by JGraph [169] (an extension of Java Swing GUI Component to support directed graphs). This platform is composed of four components, (1) the Provenance View presents a graphical view of the multiple-grained levels of the scientific investigation process modelled using RDF graphs, (2) the Publishing Interface enables users to author scientific compound objects for publishing scientific results, (3) the Web Browser is implemented using Java's JDesktop Integration Component (JDIC) [173] for searching for and importing digital objects into the publishing compound object and, (4) the metadata panel provides functionality for creating and editing the metadata for the publishing compound object and for displaying the metadata of the imported digital object.

5.2.2.1 Rendering Provenance View

The rendering of the provenance view on the top LHS of the platform is a process of retrieving provenance data from the relational database, mapping the relational instances onto RDF triples and rendering a graphical view using the triples.

5.2.2.1.1 Mapping Relational Instances onto RDF Triples

In the upper corner of the LHS on Figure 5.1, there is the Provenance database that contains the experimental provenance information that is captured and stored by the WSBPEL workflow management system described in Section 2.5. Through the Semantic Layer that consists of the D2R MAP and the D2R processor, the provenance data in the forms of relational instances are mapped onto RDF triples and stored in the SWRL.OWL file. Figure 5.3 demonstrates the steps for the conversion.

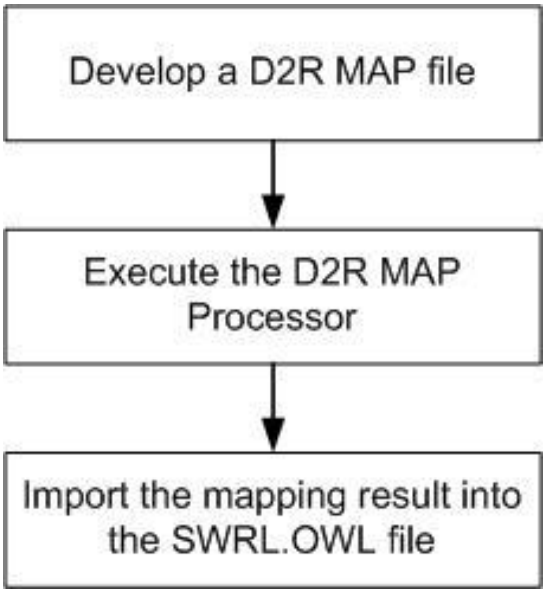


Figure 5.3: Mapping Steps

1. A D2R Map file is developed according to the MatOnto ontology described in Chapter 3. D2R MAP is a declarative language to describe the mappings between the entities and ontological classes and between the attributes and data properties. It also specifies the object properties between the classes. Figure 5.4 briefly demonstrates an extract of the D2R MAP file where the classes and object property in blue correspond to the top components on the coarse-grained view of a visualized workflow at the LHS of Figure 5.12. The D2R Map file is in Appendix C.

```
<d2r:Map xmlns:d2r="http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RMap/0.1#">
        <d2r:ClassMap
            d2r:type="comp_syn:powder"
             d2r:sql="SELECT mixing.sample_id, mixing.batch_num, powder
            FROM mixing, mixing_powder
            WHERE mixing.sample_id=mixing_powder.sample_id and
                mixing.sample_id='Melox2A-1475-4h'">
            <d2r:ObjectPropertyBridge
                d2r:property="comp_syn:input_to_synthesis"
                d2r:referredClass="comp_syn:synthesis" />
    </d2r:ClassMap>
/d2r:Map>
```

Figure 5.4: A D2R MAP Simplified Example

2. The D2R MAP Processor is executed with the D2R MAP file as input. It implements the D2R mapping language and exports data from a relational database as RDF, N3, N-TRIPLES or as JENA's models [164]. In this case, it results in an RDF/XML file.
3. The resulting file is imported into the SWRL.OWL file using JENA'S ontology package. Figure 5.5 demonstrates an extract of the file.

```
<rdf:Description rdf:nodeID="A0">
        <rdf:type rdf:resource="comp_syn:powder"/>
        <comp_syn:input_to_SB rdf:nodeID="A1"/>
</rdf:Description>
<rdf:Description rdf:nodeID="A1">
      <rdf:type rdf:resource="comp_syn:synthesis"/>
</rdf:Description>
```

Figure 5.5: The D2R Mapping Results in RDF/XML

Additionally, the SWRL.OWL file can also contain the RDF triples captured by the RDF-based e-laboratory notebooks and workflow systems, including Recentries' Collaborative Electronic Research Framework (CERF) [174], SmartTea [175] and MyTea [176], in addition to Kepler [177], Taverna [178] and Triana [179].

5.2.2.1.2 Rendering a Graphical View

The SWRL.OWL file is the input to the Provenance View of the Authoring and Publishing Platform in the RHS of Figure 5.1. A multiple-grained level graphical view of the scientific investigation workflow is rendered using the mapped RDF triples. Figure 5.6 demonstrates the rendering steps.

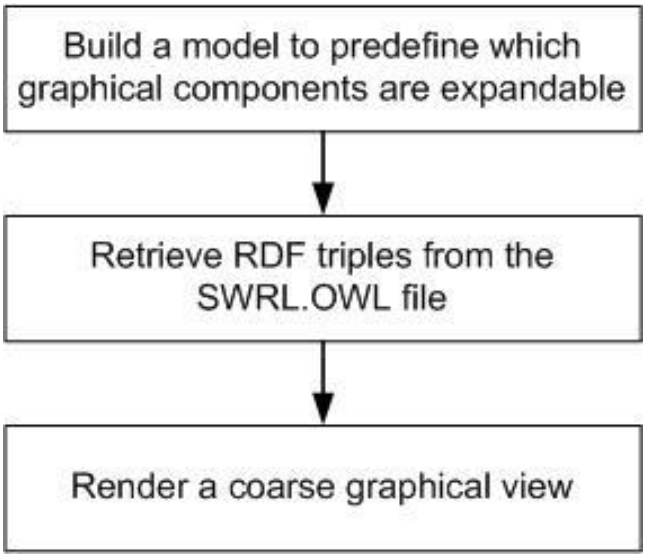


Figure 5.6: Rendering Steps

1. To develop a multiple-grained level view of a complex scientific workflow, a hash-table model is predefined using Java's HashTable class. The key element represents the expandable component while the associated object represents the expanded graph in a serialised format, for example. an RDF/XML file.

2. The system retrieves RDF triples from the SWRL.OWL file using the query package of JENA's APIs.
3. It renders a graphical view using those triples with the JGraph APIs. JGraph is an extension of the Java Swing GUI Component to support the directed graph. The initial graph is coarse, because the system just makes part of the graph visible, including the non-expandable and expandable components. However, the expanded graphs associated with the expandable components are hidden. The retrieval of further fine-grained views depends on the users' access privileges and access policies.

5.2.2.1.3 Access Controls for Fine-grained View

XACML complements Shibboleth to provide fine-grained access control on the resources. XACML, the Extensible Access Control Markup Language, provides a vocabulary for expressing the rules needed to define fine-grained and machine-readable policies and makes authorization decisions. In this system we use Sun's XACML implementation [180] that includes an XACML engine and the APIs for easy integration. Figure 5.7 demonstrates the steps for imposing the access control.

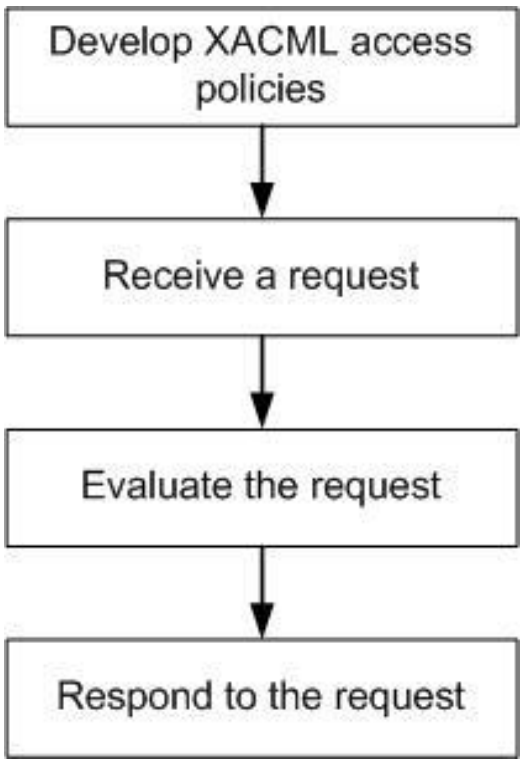


Figure 5.7: Access Control Steps

1. XACML access policies are pre-defined for every expandable component. A policy is expressed through a set of rules pertaining to a particular resource and subject. Those rules are related to whether a specific operation is permitted by a particular Subject. The Subject is described by a

set of attributes that identify the credentials of a particular user. Figure 5.8 demonstrates the policy.

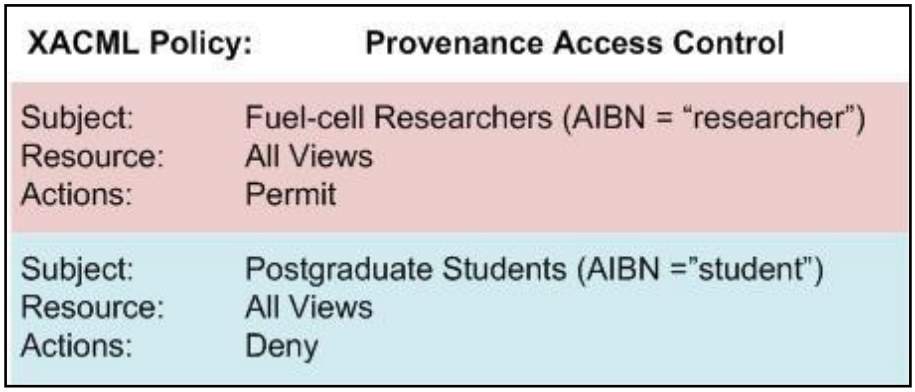


Figure 5.8: An XACML Policy

2. Initially, a user authenticated through Shibboleth is presented with the coarsest view of the provenance. When the user attempts to retrieve finer-grained views by clicking on the expandable components, a request is generated and formatted for the system evaluation by the *context schema* package of Sun's XACML Implementation APIs. The request is composed of attributes associated with the requesting user, the resource being acted upon, the action being performed on the resource and environment information.
3. The XACML engine locates the appropriate policies, compares the request with them and makes the authorization decision. This evaluation process is carried out by the core package of the XACML APIs.
4. The engine responds to the request in one of four specific types — Permit, Deny, Not Applicable or Intermediate. This response is also formatted by the *context schema* package of the XACML APIs.

5.2.2.2 Authoring

Users can author a publishing scientific compound object through the Publishing Interface on the LHS bottom of the platform. Users are able to drag and drop selected publishing components from the Provenance View into the Publishing interface as part of the publishing object. Users are also allowed to import publishing components from the Web. In many situations, scientists may want to incorporate links to external resources accessible via the Web within the compound object. Users can search for the required digital objects, such as scholarly publications and peer-reviewed datasets, through the embedded Web Browser at the RHS top of the platform and import the objects by clicking an IMPORT button at the bottom of the Browser. This adds a new node, representing the object, to the Publishing interface. The red arrow between the Web Browser and the Publishing

interface (shown in Figure 5.1) indicates the import route and the red nodes on the interface are the representations of imported web information.

In addition, within the publishing interface, any two nodes can be linked manually; then, the new direct relationships can be inferred automatically using the SWRL rules and Algernon, the reasoning engine, shown on the LHS bottom of Figure 5.1. Figure 5.9 demonstrates the inferencing steps.

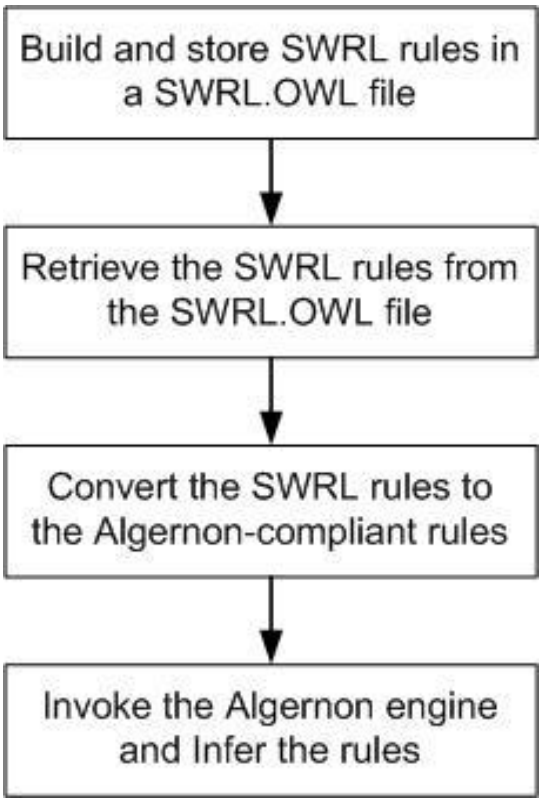


Figure 5.9: Inferencing Steps

1. SWRL rules are developed manually through Protégé shown on Figure 5.10 and stored in the SWRL.OWL file.

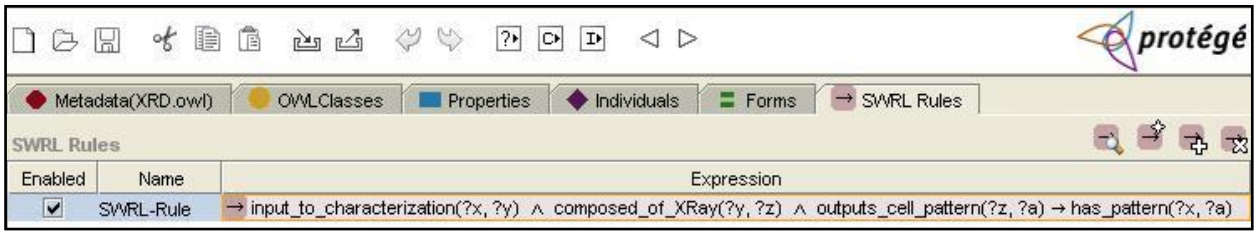


Figure 5.10: The SWRL Rule Entry in Protégé

2. Through the Protégé-OWL APIs [181], the SWRL rules are retrieved from the SWRL.OWL file for Algernon to process at runtime. Algernon is a rule-inference engine that supports both forward and backward chaining rules of inference, and implements Access-Limited Logic.

However Algernon does not support the inference of ‘subsumption’ between properties or comply with the SWRL rule format.

3. The retrieved SWRL rules are converted to the Algernon-compliant rules before being imported to Algernon at runtime.
4. The Algernon engine infers the relationship with the SWRL rules using the Algernon APIs.

5.2.2.3 Publishing

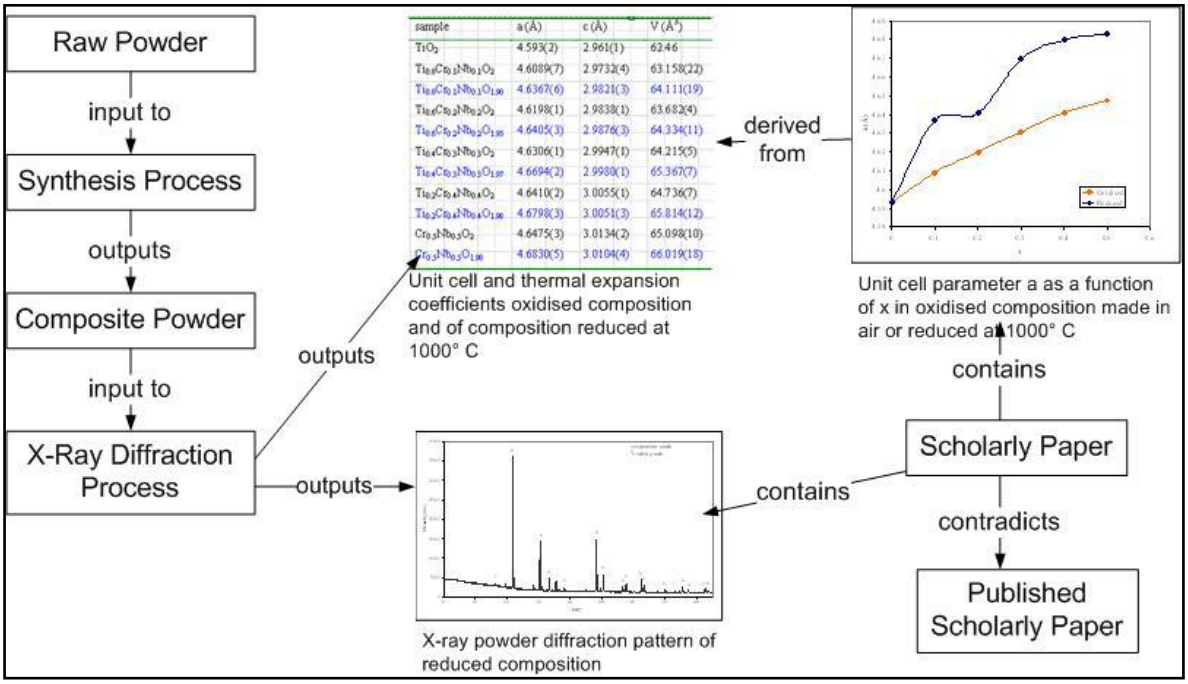
The nodes and arcs displayed in the Publishing Interface in Figure 5.1 represent the current components of the compound object that is being constructed. Before the publication, users are required to enter metadata (creator, date, title, description). In addition, users can choose a Creative Commons license and attach it to the compound object. Finally, the SCOPE system enables users to publish the object in two different ways, (1) the system converts the compound object into the OAI-ORE Resource Map and saves it in different standardised web publishing formats including ATOM 1.0, RDF/XML and RDFa (using the FORESITE’s Java Library [182]), (2) the system converts the compound object into a FOXML [183] file and ingests it into a Fedora [127] repository using JGraph2FOXML, a Java API developed by the author, and Fedora Access and Management Web Services [184].

5.3 Case Study

This case study is discussed generally in Section 1.3. This section presents a specific part of the case study details to evaluate the SCOPE system.

The composite powder that is the output from a complex synthesising process is characterized by using X-ray diffraction techniques. In addition, thermal expansion coefficients and electronic conductivity are measured over a range of operating temperatures. During the characterization process, significant amounts of data are generated in a range of formats including images, numerical data and graphs. Figure 5.11 provides a simplified view of the powder synthesis and characterization process. It also shows a subset of the datasets generated from X-Ray diffraction and characterization — that are processed to generate the graphs included in the final publication. For illustrative purposes, we also assume that the publication contradicts earlier results published in another previous publication.

The challenge is to provide a system that enables the fuel cell scientist to quickly and easily package up the relevant datasets, images, graphs and papers into a publishable compound object that also contains an explanation of the relationships between the components, the method of derivation, and allows an easy fine-grained discovery of the components.



Figures 5.11: Simplification of the Scientific Discovery Process for Novel Oxide Conductors

5.4 Implementation and User Interface

In this section, we discuss the implementation of the SCOPE system, in the context of the above case study. We assume that the experimental steps and digital objects generated during the Synthesis Process and X-Ray Diffraction are captured and stored into the distributed relational databases by the WSBPEL workflow management system discussed in Section 2.5.

5.4.1 Authoring

After a user logs onto the system as a researcher and is authenticated, the user is presented with a simple SQL search interface that enables the search and retrieval of existing relational instances and conversion of them into RDF instances through the semantic layer discussed in Section 5.2.2.1. For example, a user can search for and retrieve a particular experiment via a unique ID, for instance, EXP280818. Initially a user is presented with the basic default view of the experiment provenance in the Provenance View window shown on the LHS of Figure 5.12. The blue nodes indicate the *synthesis* and *characterization* processes that can be expanded to reveal further fine-grained information represented by the light gray nodes shown on the RHS of Figure 5.12.

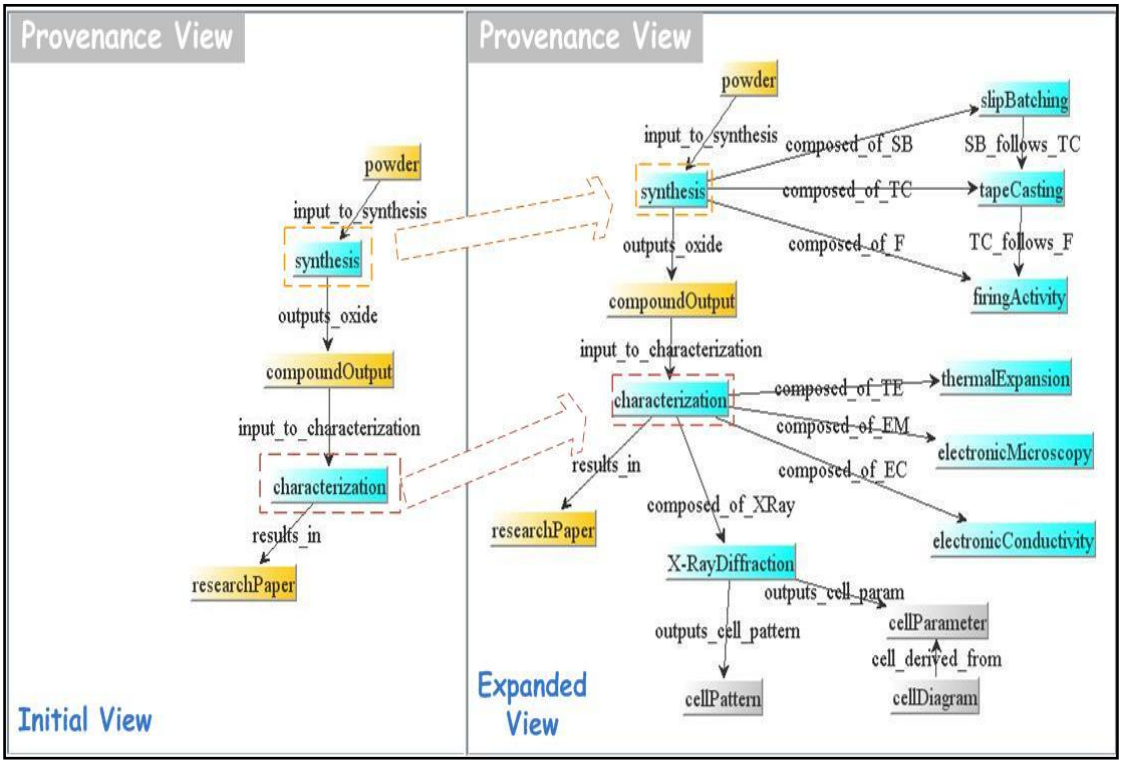


Figure 5.12: The Coarse-grained and Fine-grained Views of a Simplified Scientific Process

When the researcher clicks on a blue node, a request for additional information is generated and submitted to the XACML engine. The XACML engine compares the request with the policy and makes an authorization decision accordingly. Figure 5.13 demonstrates the policy and request.

Fuel-Cell Researchers (AIBN =”researcher”) Read All Views= “Permit”	Subject AIBN=”researcher” Resource Resource=”http://www.owl-ontologies.com/EM_ScientificProcess.ow/#SBViews” Action Action-id =”read”

Figure 5.13: Example policies and requests

By interactively drilling down via the blue nodes, the researcher is presented with the fine-grained view. The RHS of Figure 5.12 illustrates the fine-grained view. The expanded nodes including the blues (expandable) and grey (non-expandable) nodes can be collapsed manually back to the original view. A user is also allowed to examine the provenance information of every node displayed on the Provenance View. Figure 5.14 demonstrates the provenance information of node *powder*, which is shown on the Metadata Panel of the RHS bottom.

When a user starts to author a compound object, they are enabled to, (1) drag and drop the nodes up from the Provenance View down to the Publishing Interface, for example, nodes *compoundOutput* and *cellPattern,* as shown in Figure 5.15, (2) link the nodes together manually and, (3) infer the new direct relationship between them with the rule shown on the RHS of Figure 5.15 via the Algernon. The LHS top view of Figure 5.15 displays the (blue) inferencing route beginning with node *compoundOutput* and ending with node *cellPattern*, while the bottom panel shows the (circled) inferencing result —arc *has_pattern*.

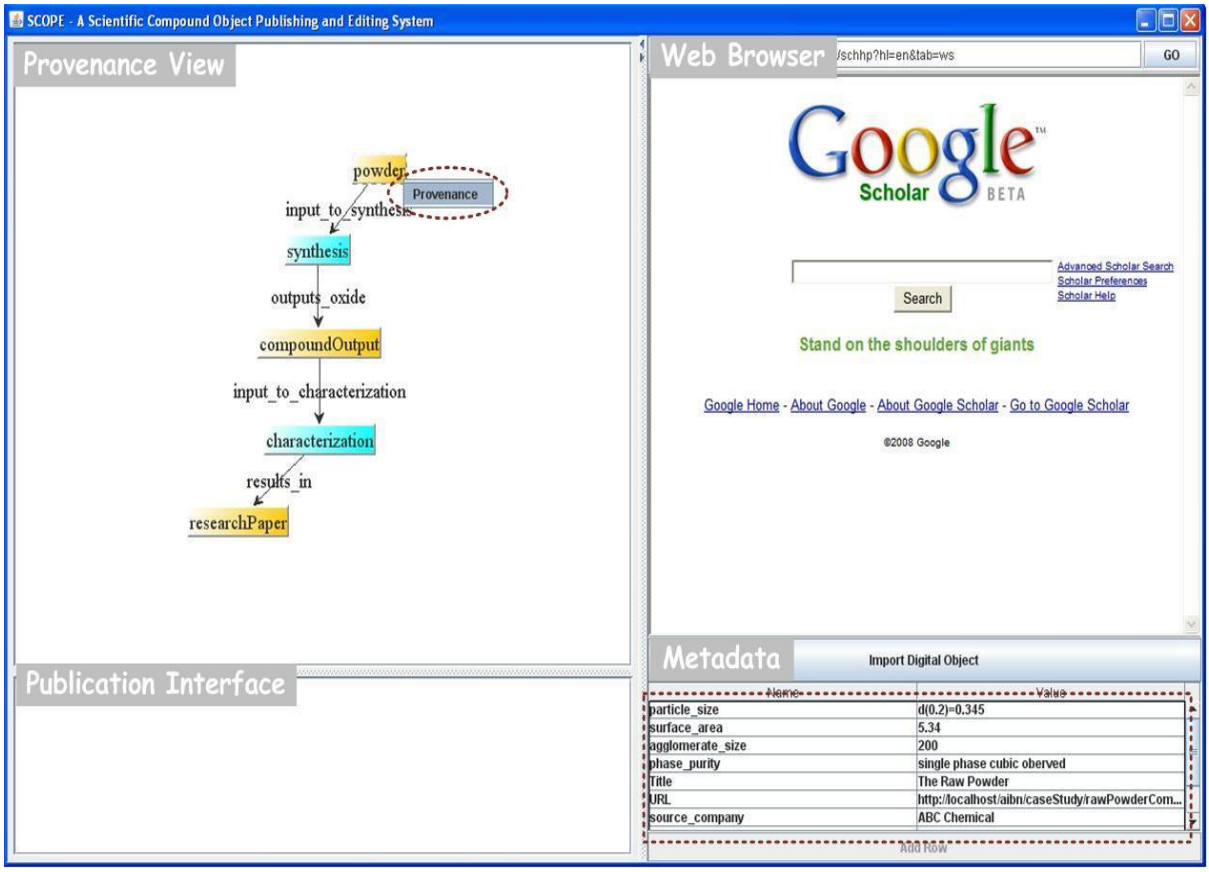


Figure 5.14: Provenance Information

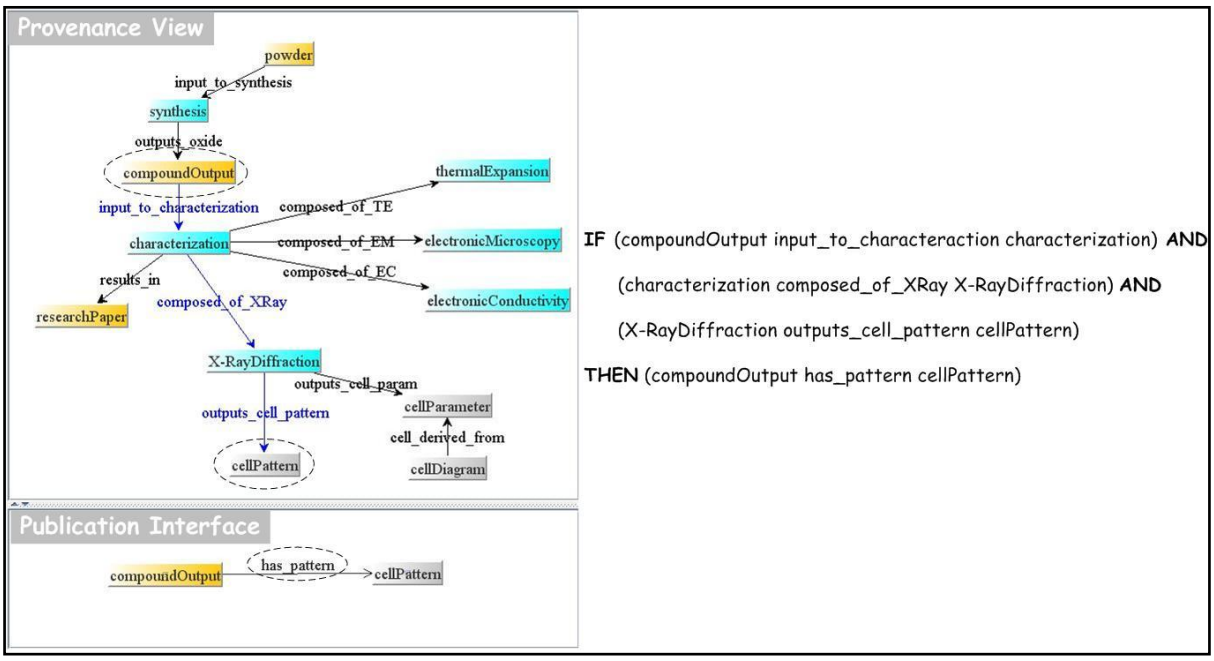


Figure 5.15: Inferencing Route and Rule

Apart from publishing the digital objects from the Provenance View, a user is allowed to search for and import digital objects discoverable via the Web through the embedded Web Browser on the RHS top of the SCOPE system. Figure 5.16 demonstrates how a scholarly paper with the metadata from the *Acta Crystallographica Section A* [185], an online journal, is imported to the Publishing Interface. The red arrow between the Publishing Interface and the Web Browser reveals the importing route. The red node *publishedPaper* results from the import, while the RHS Metadata panel displays the associated imported metadata.

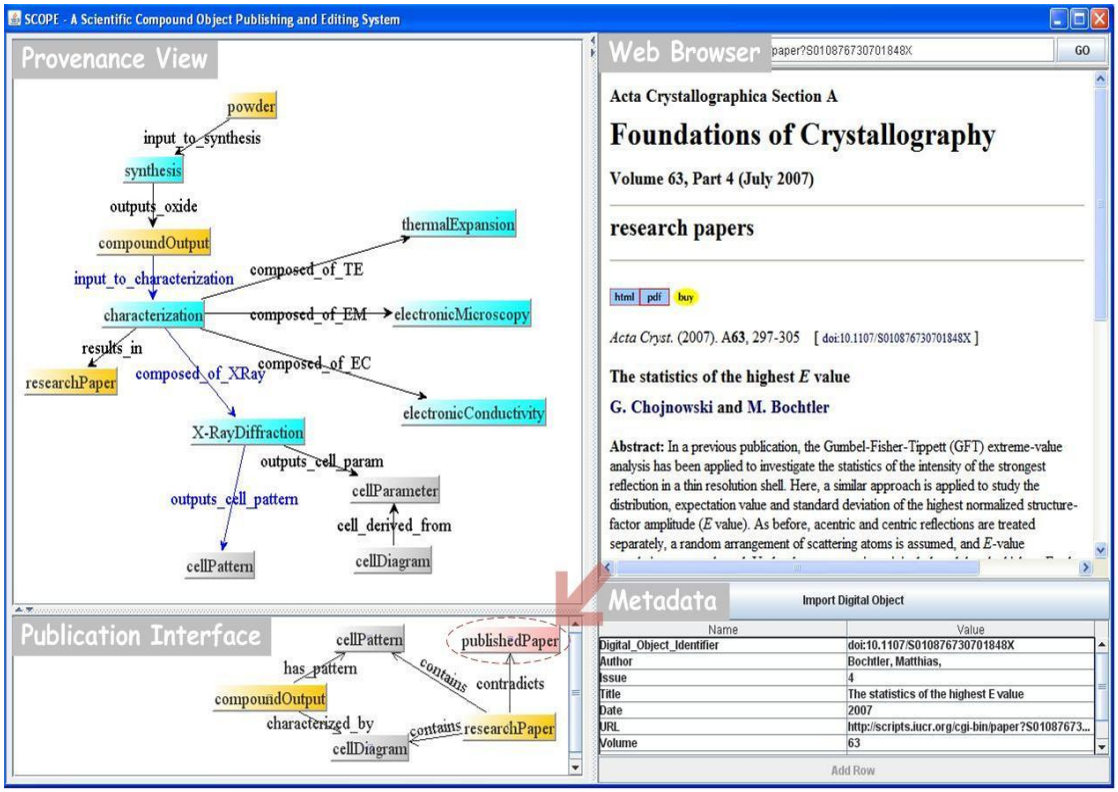


Figure 5.16: Importing Digital Objects through the Embedded Browser

5.4.2 Publishing

After completing the authoring process, a user is allowed to create and attach metadata for the new publishing compound object. Figure 5.17 demonstrates the graphical view of the publishing compound object (LHS) with the metadata (RHS). Additionally, a user is able to choose and attach one of the Common Creative licenses to the compound object shown on Figure 5.18.

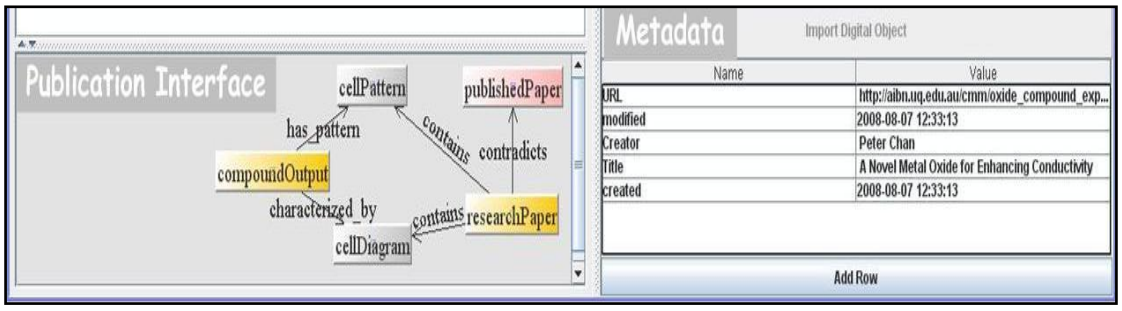


Figure 5.17: A Publishing Compound Object with the Metadata

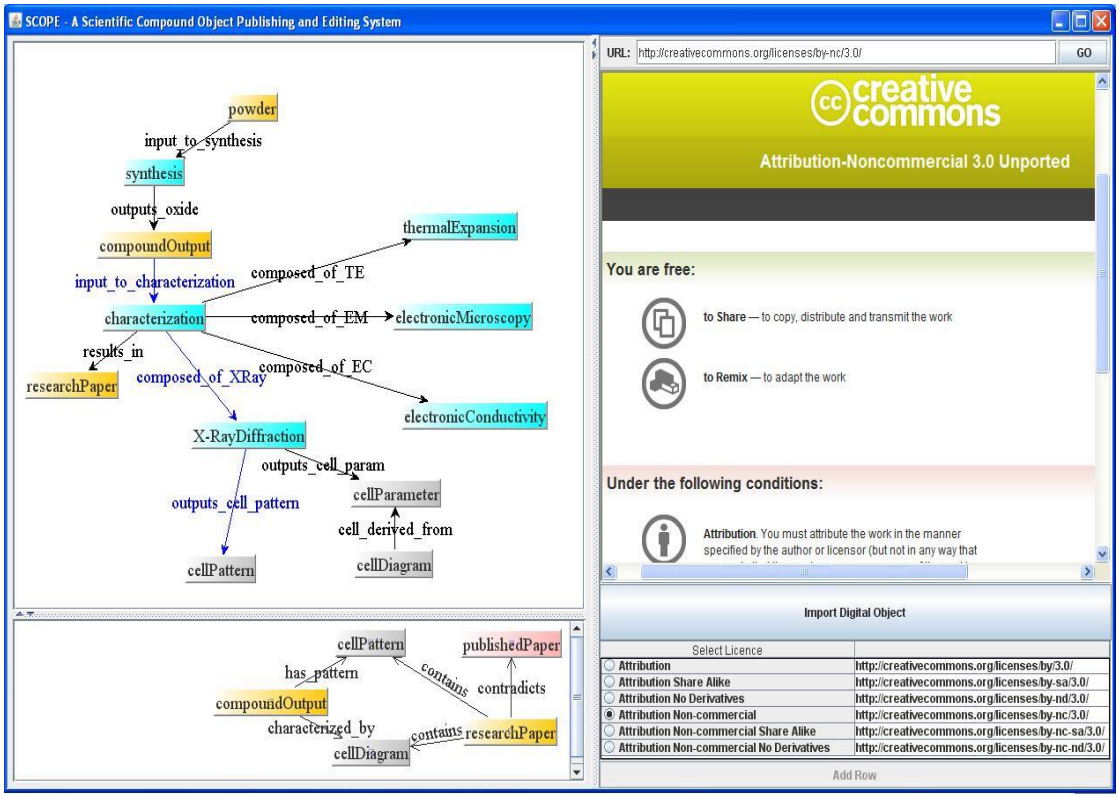


Figure 5.18: The Creative Commons License

Finally, a user is allowed to convert the visualized compound object to the OAI-ORE Resource Maps in ATOM 1.0 [186], RDF/XML [187] and RDFa [188] (shown in Figure 5.19), and convert the object to the Fedora Object XML FOXML [183] and ingest it to a Fedora Digital Library, as shown in Figure 5.20.

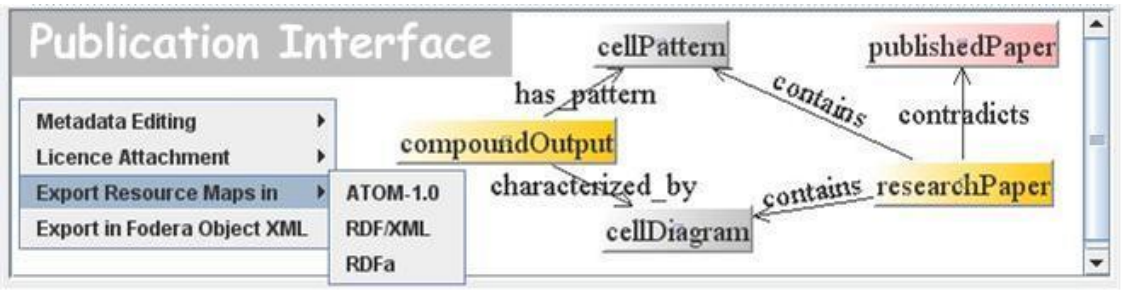


Figure 5.19: Converting to the OAI-ORE Resource Map in the formats of ATOM 1.0, RDF/XML and RDFa

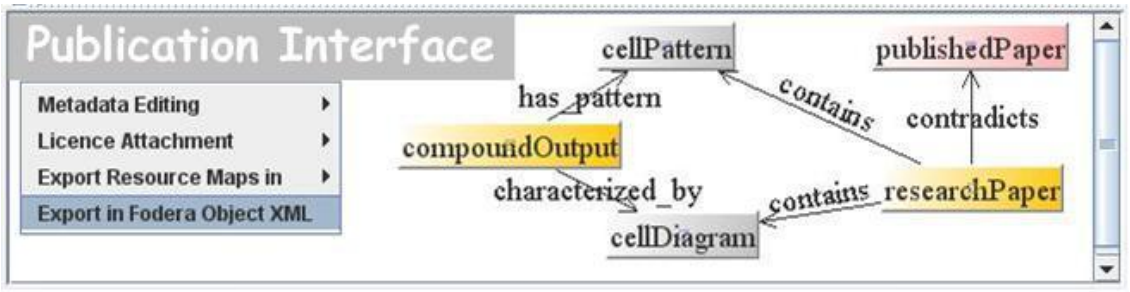


Figure 5.20: Converting to the FOXML ingested to a Fedora Digital Library

5.5 Discussion

5.5.1 User Feedback

Feedback from the collaborating scientists within AIBN was very positive. They were impressed by the novel way in which they locate and access targeted data and information through the intuitive view of the scientific experiment process. Additionally, they are able to quickly and intuitively understand quite complex workflows, so that they can easily review the methodology and justify the research outcome. They particularly liked the ability for graphically linking internally-generated provenance trails to external resources discoverable via a Web browser. This allows the authors to include other relevant research outcomes to strengthen the claims of their findings, thereby making their research outcomes more comprehensive, but still self-contained — facilitating the peer review process. The ability for interactively generating the coarse-grained view of the scientific experimental process, via automatic inferencing, was also very popular, as was being able to attach attributions and the Creative Commons license. Further collaboration with the scientists is required to develop better, discipline-specific inferencing rules and relationship ontologies, grounded in solid scientific research and knowledge.

5.5.2 Limitations and Future Work

The prototype of the SCOPE system has confirmed the usefulness of overcoming those barriers identified in Section 1.2.2. However, a number of limitations have been identified:

- Currently, only the Dublin Core metadata is supported for the FOXML files at the time of publication. Support for other metadata schemas should be an option.
- New typed relationships defined through the Publishing Interface can currently only be labelled with free text labels. We are working on an ontology that defined a class hierarchy for relationships between information objects within the scientific domain. We are also focusing more effort on the inferencing rules that apply to these relationships.

- The system currently only supports uni-directional relationships. We would like the ability to define bi-directional relationships — and symmetric, transitive and reflexive relationships within the relationship ontology.
- The expandable and collapsible functionality of the graphical nodes is hard-coded. Named Graphs [120] could be an ideal data model for this functionality. A Named Graph is a set of triples named by an URI. Thus, the expandable node/arc can be represented as a named graph, within which the expanded graph that is represented in RDF is a set of triples.
- Some of the script behaviours on the web pages could prove frustrating to users. For example, clicking on a hyperlink within the Web Browser may trigger the launch of a new browser window outside the system.
- At this stage, the system does not support searching, reloading and editing of published OAI-ORE scientific compound objects. This capacity has been planned for the future development.

Future plans also include discussing possible deployment of the SCOPE system within the Materials Science domain, in conjunction with the NSDL Materials Digital Library [189]— to enable publishing of compound scientific publication or e-learning resources on materials science research. Because this tool is not restricted to the materials science domain, we plan to evaluate the system more thoroughly using case studies and user groups from other disciplines, such as Bioinformatics, Earth Sciences, Crystallography and the Humanities and Social Sciences.

5.6 Summary

In this chapter, we have described the SCOPE system that authors and publishes OAI-ORE compliant scientific compound objects. SCOPE provides users with the hypermedia user interface that displays the graphical view of the scientific experiment process whose provenance instances have been stored in the relational and/or RDF-triple storage systems. Through the user interface, users are allowed to drill down from simple high-level views to fine-grained views of complex sub-activities by enabling graphical components to be expanded or collapsed. SCOPE has the following list of capabilities, it:

- enables users to interactively develop coarse-views scientific process for publication via automatic inferencing
- allows users to incorporate objects discoverable via the Web through an embedded Web browser
- enables users to choose and attach Creative Commons Licenses to the compound objects

- enables users to represent compound objects as the OAI-ORE Resource Maps that can be saved in ATOM 1.0, RDF/XML and RDFa
- also enables compound objects to be published as Fedora Object XML (FOXML) files within a Fedora digital library.

In delivering these capabilities, SCOPE:

- provides solutions to some of the current barriers to scientific data publishing discussed in Section 1.2.2
- provides a simple tool by which scientists can author and publish scientific compound publications that encapsulate raw data, derived data, provenance and publications in a single package.

Authors can also attach metadata to the individual components and the compound object and save the package in a variety of formats, to maximize the discovery, dissemination and re-use of the publication or its components. With the worldwide efforts for open access to publicly funded research, scientists are under increasing pressure from funding agencies to publish the experimental and evidential data with the related traditional scholarly publication (s). SCOPE can help facilitate this.

Chapter 6

6.0 Conclusion

This project has evaluated, extended and combined the emerging technologies of the Semantic Web, Web Service Business Process Execution Language (WSBPEL), and Open Archives Initiative Protocol – Object Reuse and Exchange (OAI-ORE) implement the prototype systems of the major components of the proposed Materials Informatics Workbench. These components aim at resolving the challenges confronting the materials scientists in materials science data assimilation and dissemination. They include the Materials Science Ontology (MatOnto), the ontology-based federated search interface MatSeek, the WSBPEL workflow management system and the Scientific Compound Object Publishing and Editing system (SCOPE).

6.1 Results and Significance

MatOnto has been developed and represented in OWL. MatOnto is an extensible ontology, based on the DOLCE upper ontology, that plans to represent the structured knowledge about materials, their structure and properties and the processing steps involved in their composition and engineering. MatOnto is also an underlying data model of the Materials Informatics Workbench. As a result, MatOnto enables the other major workbench components to, (1) map between and integrate disparate materials science databases, (2) model experimental provenance information captured in the physical and digital domain and, (3) inference and extract new knowledge within the materials science domain.

The MatSeek system — an ontology-based federated search interface to the key materials science database and analytical tools — has been prototyped using the Semantic Web and Web 2.0 technologies. MatSeek provides the materials scientists with a single Web interface that enables them to: search across disparate databases containing crystal structure data, ionic conductivity data and phase stability data; render 3D crystal structure images; calculate bond lengths and angles; retrieve relevant scholarly reference; and identify potential new materials with the structure and properties to satisfy specific applications. The MatOnto ontology underpinning MatSeek enables integration of data across disparate databases. The Referential Relationship Ontology enables MatSeek to search with controlled keywords, thereby resulting in an intuitive, Google-like and user-friendly search interface. The Web 2.0 technologies enable iterative searching across the databases — the retrieved results from searching the previous database are used as input to the query on the next database. The AIBN scientists were impressed with its convenience, and high efficiency and accuracy in data integration.

This project has also prototyped a Scientific Compound Object Publishing and Editing system (SCOPE) using the emerging technologies and specifications of Semantic Web and OAI-ORE. The SCOPE system is designed to enable scientists to intuitively access experimental data and information, and easily author, publish and edit scientific compound objects. First, the SCOPE system dynamically generated customized views of scientific data provenance that depend on the viewer's requirements and/or access privileges. Using RDF and graph visualization, it enables scientists to view the data, states and events associated with a scientific workflow to understand the scientific methodology and validate the results. Initially, the viewers are presented with a high-level view of a highly complex workflow; and then, expanding the view further for more details depends on their access privileges. As the result of the multiple-grained level of view approach, viewers can zoom in and out the workflow at will, while the confidential information can be guarded from unauthorised viewers. Second, as the users author a compound object, SCOPE enables them to not only select particular nodes within the visualized workflow and drag and drop them into the publishing platform, but also to incorporate objects discoverable via the Web. This allows authors to include other relevant research outcomes to strengthen the claims of their findings, thereby making their research results more comprehensive but still self-contained and facilitating the peer-review process. Third, the direct relationships between the publishing components can be inferred by a rule-inference engine. Fourth, the SCOPE system enables scientists to attach attribution in the Dublin Core format and the Creative Commons license for attribution and IP protection, respectively. Finally, the SCOPE system enables scientists to: (1) publish scientific datasets as a scientific compound object complying with OAI-ORE in a variety of standardised web publishing formats that maximise the dissemination of the compound object and, (2) export scientific datasets into a Fedora digital library directly.

6.2 Limitations and Possible Improvements

Currently, MatOnto has only been used by the AIBN scientists. It has been submitted to and reviewed by the materials science community. The initial feedback is positive. However, MatOnto's potential could not be realised unless it is acknowledged and endorsed by the community. Hopefully, engaging more community members in MatOnto's future development will help reach its greater consensus within the community.

The MatSeek system developed to date is a working prototype that demonstrates the benefits of a single entry point to the key materials databases and analysis tools. However, a number of limitations have been identified. First, there is lack of data in the publicly-available databases that we have incorporated. Commercial databases are more complete and comprehensive, but outside the scope and budget of this project. Hopefully over time, the culture of sharing materials science data

through open-access archives will become more widely adopted in the materials science community and this situation should improve. Next, the current available tools are mainly for the analysis of crystal structure data. There are many other analytical and modelling services that could be usefully incorporated. Finally, adding new databases requires human effort to map the databases schemas onto the underlying ontology and populate the names of databases, entities and attributes as instances into the ontology from those schemas. Ideally, the uploading and mapping of new databases schemas could be streamlined via a web interface.

The prototype of the SCOPE system has confirmed the usefulness and effectiveness of overcoming those barriers identified in Section 1.2.2. However, a number of limitations have been reported. First, the system lacks the functionality for searching, reloading and editing of published OAI-ORE scientific compound objects. Second, new typed relationships defined through the Publishing Interface can, currently, only be named with free text labels. An ontology is required to define a class hierarchy for relationships between information objects within the scientific domain. A set of inferencing rules that apply to these relationships is also required. Third, the system currently supports only uni-directional relationships. The ability for defining bi-directional relationships — symmetric, transitive and reflexive relationships — within the relationship ontology is required. Fourth, currently, only the Dublin Core metadata is supported for the Fedora Object XML FOXML files at the time of publication. Support for other metadata schemas should be an option. Fifth, because this system is able to ingest OAI-ORE scientific compound objects into a Fedora digital library, we plan to discuss the possible deployment of this system within the Materials Science domain, in conjunction with the NSDL Materials Digital Library ,to enable the publication of compound scientific publication or e-learning resources on materials science research. Finally, because this tool is not restricted to the materials science domain, we plan to evaluate the system more thoroughly, using case studies and user groups from other disciplines, such as Bioinformatics, Earth Sciences, Crystallography and the Humanities and Social Sciences.

References

1. A. Belsky, et al., “New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design,” *Acta Crystallographica Section B*, vol. 58, no. 3 Part 1, 2002, pp. 364 - 369; DOI:10.1107/S0108768102006948.
2. C.W. Bale, et al., “FactSage thermochemical software and databases,” *Calphad*, vol. 26, no. 2, 2002, pp. 189-228.
3. X. Yibin, et al., “Development of NIMS Materials Database and Its Application System,” *Thermophys Prop*, vol. 26, 2005, pp. 84-86.
4. F. Allen, “The Cambridge Structural Database: a quarter of a million crystal structures and rising,” *Acta Crystallographica Section B*, vol. 58, no. 3 Part 1, 2002, pp. 380-388; DOI:10.1107/S0108768102003890.
5. L.M. Bartolo, et al., “MatDL: Integrating Digital Libraries into Scientific Practice,” *Journal of Digital Information*, vol. 5, no. 3, 2004.
6. “NIST Phase Equilibria Diagrams Database,” [cited 14 Jan 2008]; Available from: http://www.ceramics.org/publications/phase.aspx.
7. “Nature - Editorial Policies about Availability of Data and Materials,” [cited 28 Aug 2007]; Available from: http://www.nature.com/authors/editorial_policies/availability.html.
8. “American Chemical Society,” [cited 13 June 2008]; Available from: http://portal.acs.org/portal/acs/corg/content.
9. “Acta Crystallographica,” [cited 27 Aug 2007]; Available from: http://journals.iucr.org/.
10. P. Murray-Rust, “Data-driven Science - A Scientist's View,” *Proc. NSF/JISC Repositories Workshop*, 2007.
11. G. Cooper, “Generalizations in Ecology: A Philosophical Taxonomy,” *Biology and Philosophy*, vol. 13, no. 4, 1998, pp. 555-586.
12. G. Shiflet, “The More Elements, the Merrier. (Materials Science),” *Science*, vol. 300, no. 5618, 2003, pp. 443-444.
13. A. Franceschetti and A. Zunger, “The Inverse Band-Structure Problem of Finding an Atomic Configuration with Given Electronic Properties,” *Nature*, vol. 402, no. 6757, 1999, pp. 60-63.
14. T. Wang, et al., “Deliberately Designed Materials for Optoelectronics Applications,” *Physical Review Letters*, vol. 82, no. 16, 1999, pp. 3304.
15. K. Rajan, “Materials informatics,” *Materials Today*, vol. 8, no. 10, 2005, pp. 38-45.
16. P.M.K. Gordon, et al., “Using a Novel Data Transformation Technique to Provide the EMBOSS Software Suite as Semantic Web Services,” *Proc. Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, 2007, pp. 117-124.
17. R.D. Stevens, et al., “myGrid: Personalised Bioinformatics on the Information Grid,” *Bioinformatics*, vol. 19, no. suppl_1, 2003, pp. i302-304; DOI 10.1093/bioinformatics/btg1041.
18. J. Gray, et al., “Scientific data management in the coming decade,” *SIGMOD Rec.*, vol. 34, no. 4, 2005, pp. 34-41; DOI http://doi.acm.org/10.1145/1107499.1107503.
19. M. Lutz, et al., “Overcoming semantic heterogeneity in spatial data infrastructures,” *Comput. Geosci.*, vol. 35, no. 4, 2009, pp. 739-752; DOI http://dx.doi.org/10.1016/j.cageo.2007.09.017.
20. W. Michener, et al., “A knowledge environment for the biodiversity and ecological sciences,” *Journal of Intelligent Information Systems*, vol. 29, no. 1, 2007, pp. 111-126.
21. W. Hunt, “Materials informatics: Growing from the Bio World,” *JOM Journal of the Minerals, Metals and Materials Society*, vol. 58, no. 7, 2006, pp. 88.
22. S.J.L. Billinge, et al., “From Cyberinfrastructure to Cyberdiscovery in Materials Science: Enhancing Outcomes in Materials Research, Education and Outreach,” *Proc. Report from NSF-sponsored workshop held in Arlington, Virginia*, 2006.
23. J.D. Wren, “404 Not Found: The Stability and Persistence of URLs Published in MEDLINE,” *Bioinformatics*, vol. 20, no. 5, 2004, pp. 668-672; DOI 10.1093/bioinformatics/btg465.
24. M. Termens, “DOI: The “Big Brother” in the dissemination of scientific documentation,” *International Microbiology*, vol. 9, no. 2, 2006, pp. 139-142.
25. A. Kalyanpur, et al., “Owl: Capturing Semantic Information Using a Standardized Web Ontology Language,” *Multilingual Computing & Technology Magazine*, vol. 15, no. 7.

26. I. Horrocks, et al., “SWRL: A Semantic Web Rule Language Combining OWL and RuleML,” [cited 29 Aug 2008]; Available from: http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/.
27. E. Sirin, et al., “Pellet: A practical OWL-DL reasoner,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, 2007, pp. 51-53.
28. C. Goble, “Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics,” *Proc. Workshop on Data Derivation and Provenance*, 2002.
29. J. Klump, et al., “Data Publication in the Open Access Initiative,” *Data Science Journal*, vol. 5, 2006, pp. 79-83.
30. P.W. Arzberger, et al., “Promoting Access to Public Research Data for Scientific, Economic, and Social Development,” *Data Science Journal*, vol. 3, 2004, pp. 135-152.
31. “Final NIH Statement on Sharing Research Data,” [cited 25 Aug 2007]; Available from: http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.
32. “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities,” [cited 23 July 2008]; Available from: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.
33. J.R. Helliwell, et al., “The Role of Quality in Providing Seamless Access to Information and Data in e-Science; the Experience Gained in Crystallography,” *Inf. Serv. Use*, vol. 26, no. 1, 2006, pp. 45-55.
34. “electronic Publication Information Center (ePIC),” [cited 31 Aug 2007]; Available from: http://epic.awi.de/Publications/Pfe2007c.pdf.
35. P. Murray-Rust and H.S. Rzepa, “The Next Big Thing: From Hypermedia to Datuments,” *Journal of Digital Information*, vol. 5, no. 1, 2004.
36. “XSL Transformations (XSLT),” [cited 30 Aug 2009]; Available from: http://www.w3.org/TR/xslt.
37. H.M. Berman, et al., “The Protein Data Bank: A Case Study in Management of Community Data,” *Current Proteomics*, vol. 1, 2004, pp. 49-57.
38. S. Coles, et al., “The 'End to End' Crystallographic Experiment in an e-Science Environment: From Conception to Publication,” *Proc. UK e-Science All Hands Meeting*, 2005.
39. L. Lyon, “eBank UK: Building the Links Between Research Data, Scholarly Communication and Learning,” *Ariadne* [cited 30 August 2008] 2003; Available from: http://www.ariadne.ac.uk/issue36/lyon/intro.html.
40. “National Center for Biotechnology Information,” [cited 25 Aug 2008]; Available from: http://www.ncbi.nlm.nih.gov/.
41. “International Virtual Observatory Alliance,” [cited 31 Aug 2007]; Available from: http://www.ivoa.net/Documents/.
42. “GBIF Biodiversity Data Portal,” [cited 31 Aug 2007]; Available from: http://data.gbif.org/datasets/.
43. M.K. Bergman, “The Deep Web: Surfacing Hidden Value,” *Journal of Electronic Publishing*, vol. 7, no. 1, 2001.
44. N.Q. Minh, “Ceramic Fuel Cells,” *Journal of the American Ceramic Society*, vol. 76, no. 3, 1993, pp. 563-588; DOI:10.1111/j.1151-2916.1993.tb03645.x.
45. A. Petric, “Ceramic fuel cells: A Century of Research,” *Canadian Ceramics*, vol. 68, no. 3, 1999, pp. 63-69.
46. S.J. Skinner and J.A. Kilner, “Oxygen Ion Conductors,” *Materials Today*, vol. 6, no. 3, 2003, pp. 30-37.
47. J. Hunter, et al., “Realizing the Hydrogen Economy through Semantic Web Technologies,” *Intelligent Systems, IEEE*, vol. 19, no. 1, 2004, pp. 40-47.
48. X. Li, et al., eds., Advanced Data Mining and Applications (ADMA 2006), Lecture Notes in Computer Science - Lecture Notes in Artificial Intelligence, Springer, 2006.
49. M.S. Islam, “ Ionic Transport in ABO3 Perovskite Oxides: A Computer Modelling Tour ” *J. Mater. Chem*, vol. 10, 2000, pp. 1027 - 1038; DOI 10.1039/a908425h.
50. M.S. Islam, “Computer Modelling of Defects and Transport in Perovskite Oxides,” *Solid State Ionics*, vol. 154-155, 2002, pp. 75-85.
51. J.B. Goodenough, “Ceramic technology: Oxide-ion conductors by design,” *Nature*, vol. 404, no. 6780, 2000, pp. 821-823.
52. S.P.S. Badwal, et al., “Investigation of the stability of ceria-gadolinia electrolytes in solid oxide fuel cell environments,” *Solid State Ionics*, vol. 121, no. 1-4, 1999, pp. 253-262.
53. A.F. Sammells, et al., “Rational Selection of Advanced Solid Electrolytes for Intermediate Temperature Fuel Cells,” *Solid State Ionics*, vol. 52, no. 1-3, 1992, pp. 111-123.

54. "Web Services Business Process Execution Language Version 2.0," *OASIS Standard*, [cited 10 June 2008]; Available from: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html.
55. C. Lagoze, et al., "Object Re-Use & Exchange: A Resource-Centric Approach," [cited 29 Aug 2008]; Available from: http://arxiv.org/abs/0804.2273
56. J. Hunter, "Scientific Models - A User-oriented Approach to Integrating Scientific Data and Digital Libraries," *Proc. VALA 2006*.
57. Y. Li, "Building The Data Warehouse For Materials Selection in Mechanical Design," *Advanced Engineering Materials*, vol. 6, no. 1-2, 2004, pp. 92-95.
58. L. Maurizio, "Data Integration: A Theoretical Perspective," *Proc. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, 2002.
59. P. Murray-Rust and H.S. Rzepa, "Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles," *J. Chem. Inf. Comput. Sci.*, vol. 39, no. 6, 1999, pp. 928-942.
60. P. Murray-Rust and H.S. Rzepa, "Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM," *J. Chem. Inf. Comput. Sci.*, vol. 41, no. 5, 2001, pp. 1113-1123.
61. G.V. Gkoutos, et al., "Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust," *J. Chem. Inf. Comput. Sci.*, vol. 41, no. 5, 2001, pp. 1124-1130.
62. P. Murray-Rust and H.S. Rzepa, "Chemical Markup, XML, and the World Wide Web. 4. CML Schema," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 3, 2003, pp. 757-772.
63. P. Murray-Rust, et al., "Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 2, 2004, pp. 462-469.
64. G.L. Holliday, et al., "Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions," *J. Chem. Inf. Model.*, vol. 46, no. 1, 2006, pp. 145-157.
65. S. Kuhn, et al., "Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data," *J. Chem. Inf. Model.*, vol. 47, no. 6, 2007, pp. 2015-2034.
66. L.M. Bartolo, et al., "Use of MatML with software applications for e-learning," *Proc. Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, 2004, pp. 190-191.
67. C.P. Sturrock, et al., *MatML – Materials Markup Language Workshop Report*, National Institute of Standards and Technology, 2001.
68. R. Lowner, et al., "Field Data and the Gas Hydrate Markup Language," *Data Science Journal*, vol. 6, 2007, pp. 6-17.
69. W. Wang, et al., "Modeling Hydrates and the Gas Hydrate Markup Language," *Data Science Journal*, vol. 6, 2007, pp. 25-36.
70. P.E. van der Vet, et al., "The PLINIUS Ontology of Ceramic Materials," *Proc. the Eleventh European Conference on Artificial Intelligence (ECAI'94) Workshop on Comparison of Implemented Ontologies*, The Netherlands, 1994.
71. M. Tanaka, "Toward a Proposed Ontology for Nanoscience," *Proc. CAIS/ACSI 2005: Data, Information, and Knowledge in a Networked World*, 2005.
72. T. Ashino, "Material Ontology (version 1.1)," [cited 5 April 2008]; Available from: http://www.codata.jp:8080/doc/MaterialOntology_v1.1.pdf.
73. T. Ashino and N. Oka, "Material Ontology: an Infrastructure for exchanging material information and knowledge," *Proc. 21st International CODATA Conference*, 2008.
74. X. Zhang, et al., "Material Scientific Data Integration for Semantic Grid," *Proc. Semantics, Knowledge and Grid, Third International Conference on*, 2007, pp. 414-417.
75. "Material Data Network," [cited 5 April 2008]; Available from: http://www.matdata.net/index.jsp.
76. "MatNavi: NIMS Materials Database," [cited 5 April 2008]; Available from: http://mits.nims.go.jp/db_top_eng.htm.
77. T.R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal Human-Computer Studies*, vol. 43, no. 5-6, 1995, pp. 907-928.
78. J. Yu and R. Buyya, "A taxonomy of scientific workflow systems for grid computing," *SIGMOD Rec.*, vol. 34, no. 3, 2005, pp. 44-49; DOI http://doi.acm.org/10.1145/1084805.1084814.
79. M. Hassan, et al., "Cheminformatics analysis and learning in a data pipelining environment," *Molecular Diversity*, vol. 10, no. 3, 2006, pp. 283-299.
80. "InforSense KDE," [cited 17 July 2008]; Available from: http://www.inforsense.com/pdfs/InforSense_KDE_Factsheet.pdf.

81. M. Peeler, “Workflows and Data Pipelines,” *In Silico Technologies in Drug Target Identification and Validation* Drug Discovery Series, D. León and S. Markel, eds., CRC/Taylor & Francis, 2006, pp. 427.
82. D. Farrusseng, et al., “Development of an Integrated Informatics Toolbox: HT Kinetic and Virtual Screening,” *Combinatorial Chemistry & High Throughput Screening*, vol. 10, 2007, pp. 85-97; DOI:10.2174/138620707779940947.
83. “Towards Optimised Chemical Processes and New Materials by Combinatorial Science (TOPCOMBI),” [cited 23 July 2008]; Available from: http://www.topcombi.org/.
84. S. AlSairafi, et al., “The Design of Discovery Net: Towards Open Grid Services for Knowledge Discovery,” *International Journal of High Performance Computing Applications*, vol. 17, no. 3, 2003, pp. 297-315; DOI 10.1177/1094342003173003.
85. W.E. Dietmar, “UNICORE - A Grid Computing Environment,” *Concurrency and Computation: Practice and Experience*, vol. 14, no. 13-15, 2002, pp. 1395-1410.
86. T. Kuhn, et al., “Creating chemo- & bioinformatics workflows, further developments within the CDK-Taverna Project,” *Chemistry Central Journal*, vol. 2, no. Suppl 1, 2008, pp. P27.
87. D. Hull, et al., “Taverna: A Tool for Building and Running Workflows of Services,” *Nucl. Acids Res.*, vol. 34, no. suppl_2, 2006, pp. 729-732; DOI 10.1093/nar/gkl320.
88. C. Steinbeck, et al., “The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, 2003, pp. 493-500.
89. C. Steinbeck, et al., “Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics,” *Current Pharmaceutical Design*, vol. 12, no. 17, 2006, pp. 2111-2120.
90. O. Spjuth, et al., “Bioclipse: an open source workbench for chemo- and bioinformatics,” *BMC Bioinformatics*, vol. 8, no. 1, 2007, pp. 59.
91. T. Oinn, et al., “Delivering Web Service Coordination Capability to Users,” *Proc. WWW2004*, 2004.
92. T. Fahringer, et al., “ASKALON: A Development and Grid Computing Environment for Scientific Workflows,” *Workflows for e-Science*, 2007, pp. 450-471.
93. T. Fahringer, et al., “A-GWL: Abstract Grid Workflow Language,” *Computational Science - ICCS 2004*, 2004, pp. 42-49.
94. P. Blaha, et al., *WIEN2k: An Augmented PlaneWave Plus Local Orbitals Program for Calculating Crystal Properties*, Inst. of Physical and Theoretical Chemistry, Vienna University of Technology, 2001.
95. A. Akram, et al., “Evaluation of BPEL to Scientific Workflows,” *Proc. Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on*, 2006, pp. 269-274.
96. M.J. Duftler, et al., “Web Services Invocation Framework WSIF,” *Proc. the OOPSLA Workshop on Object–Oriented Web Services*, 2001.
97. “The ActiveBPEL Community Edition Engine,” [cited 10 June 2008]; Available from: http://www.activevos.com/community-open-source.php.
98. “Apache Orchestration Director Engine ODE,” [cited 22 Aug 2008]; Available from: http://ode.apache.org/.
99. F.N. da Silva, et al., “In Services: Data Management for In Silico Workflows,” *Proc. Database and Expert Systems Applications, 2006. DEXA '06. 17th International Conference on*, 2006, pp. 206-210.
100. A. Malinova and S. Gocheva-Ilieva, “Using the Business Process Execution Language for Managing Scientific Processes,” *Information Technologies & Knowledge*, vol. 2, 2008, pp. 257 - 261.
101. S. Aleksander, “On using BPEL extensibility to implement OGSI and WSRF Grid workflows: Research Articles,” *Concurr. Comput. : Pract. Exper.*, vol. 18, no. 10, 2006, pp. 1229-1241; DOI http://dx.doi.org/10.1002/cpe.v18:10.
102. L. Frank, “Choreography for the Grid: towards fitting BPEL to the resource framework: Research Articles,” *Concurr. Comput. : Pract. Exper.*, vol. 18, no. 10, 2006, pp. 1201-1217; DOI http://dx.doi.org/10.1002/cpe.v18:10.
103. R. Shrija and W.W. David, “Incorporating Provenance in Service Oriented Architecture,” *Proc. Next Generation Web Services Practices, 2006. NWeSP 2006. International Conference on*, 2006, pp. 33-40.
104. A. Agrawal, et al., “WS-BPEL Extension for People (BPEL4People), Version 1.0,” [cited 10 June 2008]; Available from: http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel4people/BPEL4People_v1.pdf.

105. R. Bose and J. Frew, "Composing lineage metadata with XML for custom satellite-derived data products," *Proc. Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, 2004, pp. 275-284.
106. J.D. Myers, et al., "Multi-scale science: supporting emerging practice with semantically derived provenance," *Proc. ISWC 2003 Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data*, 2003.
107. J. Zhao, et al., "Using Semantic Web Technologies for Representing E-science Provenance," *The Semantic Web – ISWC 2004*, 2004, pp. 92-106.
108. D. Quan, et al., "Haystack: A Platform for Authoring End User Semantic Web Applications," *The SemanticWeb - ISWC 2003*, 2003, pp. 738-753.
109. D. Quan, et al., "Adenine: A Metadata Programming Language," [cited 24 July 2008]; Available from: http://groups.csail.mit.edu/haystack/documents/papers/2002/sow2002-adenine.pdf.
110. J. Freire, et al., "Managing Rapidly-Evolving Scientific Workflows," *Provenance and Annotation of Data*, 2006, pp. 10-18.
111. D.A. Benson, et al., "GenBank," *Nucl. Acids Res.*, vol. 35, no. suppl_1, 2007, pp. 21-25; DOI 10.1093/nar/gkl986.
112. B. Boeckmann, et al., "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucl. Acids Res.*, vol. 31, no. 1, 2003, pp. 365-370; DOI 10.1093/nar/gkg095.
113. "NASA," [cited 25 Aug 2008]; Available from: http://www.nasa.gov/home/index.html.
114. "National Institute of Standards and Technology NIST," [cited 25 Aug 2008]; Available from: http://www.nist.gov/.
115. "Publication and Citation of Scientific Primary Data STD-DOI," [cited 25 Aug 2008]; Available from: http://www.std-doi.de/front_content.php.
116. "National Oceanic and Atomspheric Administration NOAA," [cited 25 Aug 2008]; Available from: http://www.noaa.gov/.
117. S.J. Coles, et al., "An E-Science Environment for Service Crystallography-from Submission to Dissemination," *J. Chem. Inf. Model.*, vol. 46, no. 3, 2006, pp. 1006-1016.
118. J. Frey, et al., "CombeChem: A Case Study in Provenance and Annotation Using the Semantic Web," *Provenance and Annotation of Data*, 2006, pp. 270-277.
119. J. Klump and R. Conze, "The Scientific Drilling Database (SDDB)—Data from Deep Earth Monitoring and Sounding," *Scientific Drilling*, no. 4, 2007, pp. 30-31.
120. J.C. Jeremy, et al., "Named Graphs, Provenance and Trust," *Proc. Tthe 14th international conference on World Wide Web*, ACM Press, 2005.
121. R. Khare and T. Celik, "Microformats: A Pragmatic Path to the Semantic Web," *Proc. Proceedings of the 15th international conference on World Wide Web*, ACM Press, 2006.
122. "RDFa Primer 1.0 - Embedding RDF in XHTML," [cited 31 Aug 2007]; Available from: http://www.w3.org/TR/xhtml-rdfa-primer/.
123. "Gleaning Resource Descriptions from Dialects of Languages (GRDDL)," [cited 31 Aug 2007]; Available from: http://www.w3.org/2004/01/rdxh/spec.
124. P.B. Jeremy and J.B. Matthew, "Data at Work: Supporting Sharing in Science and Engineering," *Proc. Tthe 2003 International ACM SIGGROUP Conference on Supporting Group Work*, ACM, 2003.
125. "Creative Commons," [cited 25 Aug 2008]; Available from: http://creativecommons.org/.
126. R. Sayre, "Atom: The Standard in Syndication," *Internet Computing, IEEE*, vol. 9, no. 4, 2005, pp. 71-78.
127. C. Lagoze, et al., "Fedora: An Architecture for Complex Objects and Their Relationships," *International Journal on Digital Libraries*, vol. 6, no. 2, 2006, pp. 124-138.
128. C. Lagoze and H. Van de Sompel, "Compound Information Objects: The OAI-ORE Perspective," [cited 23 July 2008]; Available from: http://www.openarchives.org/ore/documents/CompoundObjects-200705.html.
129. J. Hunter, et al., "FUSION - A Knowledge Management System for Fuel Cell Optimization," *Proc. International Conference on Solid State Ionics. Applications: Fuel Cells*, 2005, pp. 544.
130. J. Hunter and K. Cheung, "Generating eScience Workflows from Statistical Analysis of Prior Data," *Proc. APAC'05*, 2005.
131. "Apache Tomcat," [cited 15 July 2008]; Available from: http://tomcat.apache.org/.

132. “MySQL Relational Database,” [cited 15 July 2008]; Available from: http://www.mysql.com/.
133. “The MathWorks - MATLAB and Simulink for Technical Computing,” [cited 15 July 2008]; Available from: http://www.mathworks.com/.
134. “JavaMail API,” [cited 15 July 2008]; Available from: http://java.sun.com/products/javamail/.
135. “Apache Axis,” [cited 15 July 2008]; Available from: http://ws.apache.org/axis/.
136. “IBM Business Process Execution Language for Web Services JavaTM Run Time (BPWS4J),” [cited 15 July 2008]; Available from: http://www.alphaworks.ibm.com/tech/bpws4j.
137. D. De Roure and C. Goble, “myExperiment - A Web 2.0 Virtual Research Environment,” *Proc. International Workshop on Virtual Research Environments and Collaborative Work Environments*, 2007.
138. R.M. Colomb, “Formal versus Material Ontologies for Information Systems Interoperation in the Semantic Web,” *The Computer Journal*, vol. 49, no. 1, 2006, pp. 4-19; DOI 10.1093/comjnl/bxh147.
139. G. Aldo, et al., “Sweetening Ontologies with DOLCE,” *Proc. The 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, Springer-Verlag, 2002.
140. T.R. Gruber and G.R. Olsen, “An Ontology for Engineering Mathematics,” *Proc. Fourth International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann, 1994.
141. T.R. Gruber, *Ontolingua: A Mechanism to Support Portable Ontologies*, KSL-91-66, Knowledge Systems Laboratory, Stanford University, 1992.
142. “The Joint Academic Classification of Subjects,” [cited 1 November 2007]; Available from: http://www.hesa.ac.uk/dox/jacs/JACS_complete.pdf.
143. “Time Ontology in OWL - W3C Editor's Draft,” [cited 20 September 2008]; Available from: http://www.isi.edu/~pan/SWBP/time-ontology-note/time-ontology-note.html.
144. Y. Sure, et al., “The Ontology – Semantic Web for Research Communities,” *Progress in Artificial Intelligence*, 2005, pp. 218-231.
145. “Higher Education Statistics Agency HESA,” [cited 11 July 2008]; Available from: http://www.hesa.ac.uk/.
146. “Universities and Colleges Admission Service UCAS,” [cited 11 July 2008]; Available from: http://www.ucas.com/.
147. L.N. Soldatova and R.D. King, “An Ontology of Scientific Experiments,” *Journal of The Royal Society Interface*, vol. 3, no. 11, 2006, pp. 795-803.
148. C. Lagoze and J. Hunter, “The ABC Ontology and Model,” *Journal of Digital Information*, vol. 2, no. 2, 2001.
149. M. Ashby, et al., *Materials : Engineering, Science, Processing and Design*, Butterworth-Heinemann, 2007.
150. , *Springer Handbook of Materials Measurement Methods*, Springer, 2006.
151. I.D. Browna and B. McMahonb, “CIF: The Computer Language of Crystallography,” *Acta Crystallographica Section B - Structural Science*, vol. 58, 2002, pp. 317-324; DOI:10.1107/S0108768102003464
152. A. Pease, “SUMO: A Sharable Knowledge Resource with Linguistic Inter-Operability,” *Proc. Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*, 2003, pp. 827.
153. J. Hunter, “Adding Multimedia to the Semantic Web: Building an MPEG-7 Ontology,” *Proc. International Semantic Web Working Symposium (SWWS)*, 2001.
154. K. Cheung, et al., “MatSeek: An Ontology-Based Federated Search Interface for Materials Scientists,” *IEEE Intelligent Systems*, vol. 24, no. 1, 2009, pp. 47-56; DOI 10.1109/MIS.2009.13.
155. R. Shannon, “Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides,” *Acta Crystallographica Section A*, vol. 32, no. 5, 1976, pp. 751-767; DOI doi:10.1107/S0567739476001551.
156. K. Cheung, et al., “SCOPE - A Scientific Compound Object Publishing and Editing System,” *Proc. 3rd International Digital Curation Conference*, 2007.
157. K. Cheung, et al., *Towards an Ontology for Data-driven Discovery of New Materials*, SS-08-05, M. Deborah, et al., Association for the Advancement of Artificial Intelligence, 2008.
158. P. Srinath, et al., “Axis2, Middleware for Next Generation Web Services,” *Proc. Web Services, 2006. ICWS '06. International Conference on*, 2006, pp. 833-840.

159. "JavaServer Pages Technology," [cited 5 April 2008]; Available from: http://java.sun.com/products/jsp/.
160. "SPARQL JavaScript Library," [cited 5 April 2008]; Available from: http://www.thefigtrees.net/lee/sw/sparql.js.
161. "OpenLink AJAX Toolkit Framework ", [cited 5 April 2008]; Available from: http://oat.openlinksw.com/.
162. K. Smith, "Simplifying Ajax-style Web development," *Computer*, vol. 39, no. 5, 2006, pp. 98-101.
163. M.L. Reuven, "At the forge: Dojo," *Linux J.*, vol. 2007, no. 155, 2007, pp. 10.
164. B. McBride, "Jena: a semantic Web toolkit," *Internet Computing, IEEE*, vol. 6, no. 6, 2002, pp. 55-59.
165. M. Smith, et al., "OWL Web Ontology Language Guide," [cited 8 April 2008]; Available from: http://www.w3.org/TR/owl-guide/#owl_equivalentClass.
166. "The R Project for Statistical Computing," [cited 8 April 2008]; Available from http://www.r-project.org/.
167. E. Frank, et al., "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, 2004, pp. 2479-2481; DOI 10.1093/bioinformatics/bth261.
168. C. Bizer, "D2R MAP – A Database to RDF Mapping Language," *Proc. WWW 2003*, 2003.
169. "Java Graph Visualization and Layout JGraph," [cited 15 Sept 2008]; Available from: http://www.jgraph.com/.
170. M. Hewett, "Algernon - Rule-Based Programming," [cited 15 Sept 2008]; Available from: http://algernon-j.sourceforge.net/.
171. M. Needleman, "The Shibboleth Authentication/Authorization System," *Serials Review*, vol. 30, no. 3, 2004, pp. 252-253.
172. "OASIS eXtensible Access Control Markup Language (XACML)," [cited 15 Sept 2008]; Available from: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml.
173. "JDesktop Integration Component," [cited 27 Aug 2007]; Available from: https://jdic.dev.java.net/.
174. "CERF - Collaborative Electronic Research Framework," [cited 29 Aug 2007]; Available from: http://www.rescentris.com/.
175. m.c. schraefel, et al., "Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment," *Proc. The SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, 2004.
176. A. Gibson, et al., "myTea: Connecting the Web to Digital Science on the Desktop," *Proc. World Wide Web Conference 2006*, 2006.
177. I. Altintas, et al., "Provenance Collection Support in the Kepler Scientific Workflow System," *Proc. International Provenance and Annotation Workship (IPAW'06)*, 2006.
178. T. Oinn, et al., "Taverna: A Tool for the Composition and Enactment of Bioinformatics Workflows," *Bioinformatics*, vol. 20, no. 17, 2004, pp. 3045-3054; DOI 10.1093/bioinformatics/bth361.
179. S. Majithia, et al., "Triana: A Graphical Web Service Composition and Execution Toolkit," *Proc. Web Services, 2004. Proceedings. IEEE International Conference on*, 2004, pp. 514-521.
180. "Sun's XACML Implementation," [cited 15 Sept 2008]; Available from: http://sunxacml.sourceforge.net/guide.html.
181. H. Knublauch, et al., "The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications," *Proc. The International Semantic Web Conference (ISWC)*, 2004, pp. 229-243.
182. R. Sanderson, et al., "Functional Object Re-use and Exchange: Supporting Information Topology Experiments," [cited 15 Sept 2008]; Available from: http://foresite.cheshire3.org/.
183. "Introduction to Fedora Object XML (FOXML)," [cited 15 Sept 2008]; Available from: http://www.fedora.info/download/2.0/userdocs/digitalobjects/introFOXML.html.
184. "Fedora Access and Management Web Services - API Documentation," [cited 15 June 2008]; Available from: http://www.fedora.info/definitions/1/0/api/.
185. "Section A of Acta Crystallographica," [cited 15 Sept 2008]; Available from: http://ww1.iucr.org/journals/acta/actaa.html.
186. "ORE User Guide - Resource Map Implementation in Atom," [cited 30 Aug 2008]; Available from: http://www.openarchives.org/ore/0.9/atom-implementation.html.
187. "ORE User Guide - Resource Map Implementation in RDF/XML," [cited 15 Sept 2008]; Available from: http://www.openarchives.org/ore/0.9/rdfxml.html.

188. “ORE User Guide - Resource Map Implementation in RDFa,” [cited 15 Sept 2008]; Available from: http://www.openarchives.org/ore/0.9/rdfa.html.
189. L. Bartolo, et al., “NSDL MatDL: Adding Context to Bridge Materials e-Research and e-Education,” *Research and Advanced Technology for Digital Libraries*, 2007, pp. 499-500.

Appendix A: The MatOnto Ontology in Manchester OWL

Namespace: Mpeg7-2001 = <http://rhizomik.net/ontologies/2005/03/Mpeg7-2001.owl#>
Namespace: DOLCE-Lite = <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#>
Namespace: SUMO = <http://www.ontologyportal.org/translations/SUMO.owl#>
Namespace: owl2 = <http://www.w3.org/2006/12/owl2#>
Namespace: MatOnto = <http://aibn.uq.edu.au/cmm/MatOnto.owl#>
Namespace: ontology-07 = <http://swrc.ontoware.org/ontology-07.owl#>
Namespace: rdfs = <http://www.w3.org/2000/01/rdf-schema#>
Namespace: JACS_complete = <http://www.hesa.ac.uk/dox/jacs/JACS_complete.owl#>
Namespace: ontolingua = <http://ksl.stanford.edu/htw/dme/thermal-kb-tour/ontolingua.owl#>
Namespace: time = <http://www.w3.org/2006/time.owl#>
Namespace: owl2xml = <http://www.w3.org/2006/12/owl2-xml#>
Namespace: cif = <http://aibn.uq.edu.au/cmm/cif.owl#>
Namespace: EXPOApr19 = <http://www.hozo.jp/owl/EXPOApr19.owl#>
Namespace: xsd = <http://www.w3.org/2001/XMLSchema#>
Namespace: owl = <http://www.w3.org/2002/07/owl#>
Namespace: rdf = <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

Ontology: <http://aibn.uq.edu.au/cmm/MatOnto.owl>
Imports: <http://ksl.stanford.edu/htw/dme/thermal-kb-tour/ontolingua.owl>
Imports: <http://swrc.ontoware.org/ontology-07.owl>
Imports: <http://aibn.uq.edu.au/cmm/cif.owl>
Imports: <http://www.w3.org/2006/time.owl>
Imports: <http://www.ontologyportal.org/translations/SUMO.owl>
Imports: <http://www.hozo.jp/owl/EXPOApr19.owl>
Imports: <http://rhizomik.net/ontologies/2005/03/Mpeg7-2001.owl>
Imports: <http://www.hesa.ac.uk/dox/jacs/JACS_complete.owl>
Imports: <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl>
ObjectProperty: cif:hasRadius
ObjectProperty: investigates
    Domain:
        JACS_complete:materials_science
    Range:
        material
ObjectProperty: hasMeasurementData
    Domain:
        material
    Range:
        measurement_data
ObjectProperty: hasOperator

Domain:
experimental_step
Range:
operator
ObjectProperty: hasTime
Domain:
EXPOApr19:scientific_activity
Range:
time:temporal_entity
ObjectProperty: hasProcess
Domain:
material
Range:
process
ObjectProperty: hasInvestigator
Domain:
EXPOApr19:scientific_activity
Range:
investigator
ObjectProperty: carriedOutBy
Domain:
ontology-07:project
Range:
ontology-07:organization
ObjectProperty: outputs
Domain:
experimental_step,
EXPOApr19:scientific_activity
Range:
activitiy_output
ObjectProperty: inputTo
Domain:
activity_input
Range:
experimental_step,
EXPOApr19:scientific_activity
ObjectProperty: hasPart
Domain:
EXPOApr19:scientific_experiment
Range:
process
ObjectProperty: hasElement
Domain:

chemical_composition
Range:
chemical_element
ObjectProperty: composed_of
Domain:
EXPOApr19:scientific_activity
Range:
EXPOApr19:scientific_experiment
ObjectProperty: sa_follows
Domain:
EXPOApr19:scientific_activity
Range:
EXPOApr19:scientific_activity
SubPropertyOf:
follows
ObjectProperty: hasSetting
Domain:
EXPOApr19:experimental_equipment
Range:
setting_parameter
ObjectProperty: follows
ObjectProperty: hasProperty
ObjectProperty: hasType
Domain:
data
Range:
data_type
ObjectProperty: hasASequenceOf
Domain:
EXPOApr19:execution_of_experiment
Range:
experimental_step
ObjectProperty: hasExperiment
Domain:
ontology-07:project
Range:
EXPOApr19:scientific_experiment
ObjectProperty: es_follows
Domain:
experimental_step
Range:
experimental_step
SubPropertyOf:

follows
ObjectProperty: hasMaterialPropety
Domain:
material
Range:
material_property
SubPropertyOf:
hasProperty
ObjectProperty: hasFormula
Domain:
chemical_composition
Range:
structured_formula
ObjectProperty: hasUnit
Domain:
SUMO:number
Range:
ontolingua:standard_unit
ObjectProperty: hasEquipment
Domain:
experimental_step
Range:
EXPOApr19:experimental_equipment
ObjectProperty: associatedWith
Domain:
ontolingua:standard_dimension
Range:
ontolingua:standard_unit
ObjectProperty: categorisedInto
Domain:
material
Range:
family
Class: cif:ionic_parameter
SubClassOf:
data
Class: cif:ion
EquivalentTo:
chemical_element
SubClassOf:
DOLCE-Lite:physical_object
Class: DOLCE-Lite:physical_object
Class: ontology-07:student

SubClassOf:
DOLCE-Lite:social_agent
Class: ontology-07:project
SubClassOf:
DOLCE-Lite:accomplishment
Class: time:temporal_entity
SubClassOf:
DOLCE-Lite:temporal_quality
Class: biological
SubClassOf:
material_property
Class: setting_parameter
SubClassOf:
data
Class: measurement
SubClassOf:
process
Class: process
SubClassOf:
DOLCE-Lite:accomplishment
Class: chemical_element
EquivalentTo:
cif:ion
SubClassOf:
DOLCE-Lite:non-physical_object
Class: Mpeg7-2001:video
SubClassOf:
data_type
Class: EXPOApr19:experimental_equipment
SubClassOf:
DOLCE-Lite:physical_object
Class: JACS_complete:academic_discipline
SubClassOf:
DOLCE-Lite:non-physical_object
Class: chemical_composition
SubClassOf:
characterization_data
Class: property
SubClassOf:
DOLCE-Lite:non-physical_object
Class: DOLCE-Lite:social_agent
Class: DOLCE-Lite:quality
Class: acoustical

SubClassOf:
material_property
Class: electrical
SubClassOf:
material_property
Class: EXPOApr19:scientific_activity
SubClassOf:
DOLCE-Lite:accomplishment
Class: DOLCE-Lite:abstract
Class: DOLCE-Lite:society
SubClassOf:
material_property
Class: activitiy_output
SubClassOf:
DOLCE-Lite:endurant
Class: data_stream
SubClassOf:
data_type
Class: ontology-07:academic_staff
Class: cif:crystal_system
SubClassOf:
stative
Class: ontology-07:publication
SubClassOf:
DOLCE-Lite:endurant
Class: thermal
SubClassOf:
material_property
Class: magnetic
SubClassOf:
material_property
Class: family
SubClassOf:
DOLCE-Lite:non-physical_object
Class: DOLCE-Lite:temporal_quality
Class: SUMO:number
SubClassOf:
data_type
Class: cif:atom_site_property
SubClassOf:
property
Class: JACS_complete:materials_science
Class: chemical

SubClassOf:
material_property
Class: cif:crystalline
SubClassOf:
structure
Class: optical
SubClassOf:
material_property
Class: radiological
SubClassOf:
material_property
Class: experimental_step
SubClassOf:
DOLCE-Lite:accomplishment
Class: measurement_data
SubClassOf:
data
Class: DOLCE-Lite:non-physical_object
Class: DOLCE-Lite:endurant
Class: material_property_data
SubClassOf:
measurement_data
Class: material_property
SubClassOf:
property
Class: texture
SubClassOf:
data_type
Class: Mpeg7-2001:audio
SubClassOf:
data_type
Class: data
SubClassOf:
DOLCE-Lite:non-physical_object
Class: investigator
SubClassOf:
ontology-07:academic_staff
Class: structured_formula
SubClassOf:
DOLCE-Lite:non-physical_object
Class: performance_data
SubClassOf:
measurement_data

Class: ontolingua:standard_dimension
SubClassOf:
DOLCE-Lite:quality
lass: ontology-07:employee
SubClassOf:
DOLCE-Lite:social_agent
Class: modelling_data_and_simulation_data
SubClassOf:
Class: cif:crystallographic_property
SubClassOf:
property
Class: characterization_data
SubClassOf:
measurement_data
Class: glass
SubClassOf:
family
Class: Mpeg7-2001:image
SubClassOf:
data_type
Class: manufacturing
SubClassOf:
process
Class: ontology-07:organization
SubClassOf:
DOLCE-Lite:society
Class: amorphous
SubClassOf:
structure
Class: stative
SubClassOf:
DOLCE-Lite:event
Class: data_type
SubClassOf:
DOLCE-Lite:non-physical_object
Class: ceramic
SubClassOf:
family
Class: DOLCE-Lite:event
Class: cif:ionic_radius
SubClassOf:
data
Class: polymer

SubClassOf:
family
Class: operator
SubClassOf:
ontology-07:employee
Class: ontolingua:standard_unit
SubClassOf:
DOLCE-Lite:abstract
Class: material
SubClassOf:
DOLCE-Lite:physical_object
Class: structure
SubClassOf:
stative
Class: EXPOApr19:scientific_experiment
SubClassOf:
DOLCE-Lite:accomplishment
Class: metal
SubClassOf:
family
Class: activity_input
SubClassOf:
DOLCE-Lite:endurant
Class: elastomer
SubClassOf:
family
Class: DOLCE-Lite:accomplishment
Class: EXPOApr19:execution_of_experiment
Class: hybrid
SubClassOf:
family
Individual: ionic_radii.ION
Types:
cif:ion
Facts:
cif:hasRadius ionic_radii.IONIC_RADIUS
SameAs:
p_record.EL_SYMBOL
Individual: p_record.EL_SYMBOL
Types:
chemical_element
SameAs:
ionic_radii.ION

Individual: ionic_radii.IONIC_RADIUS
  Types:
    cif:ionic_radius

Appendix B: Simplified Compound Synthesis Workflow

Figure A.1 demonstrates the high view in Business Process Modeling Notation BPMN.

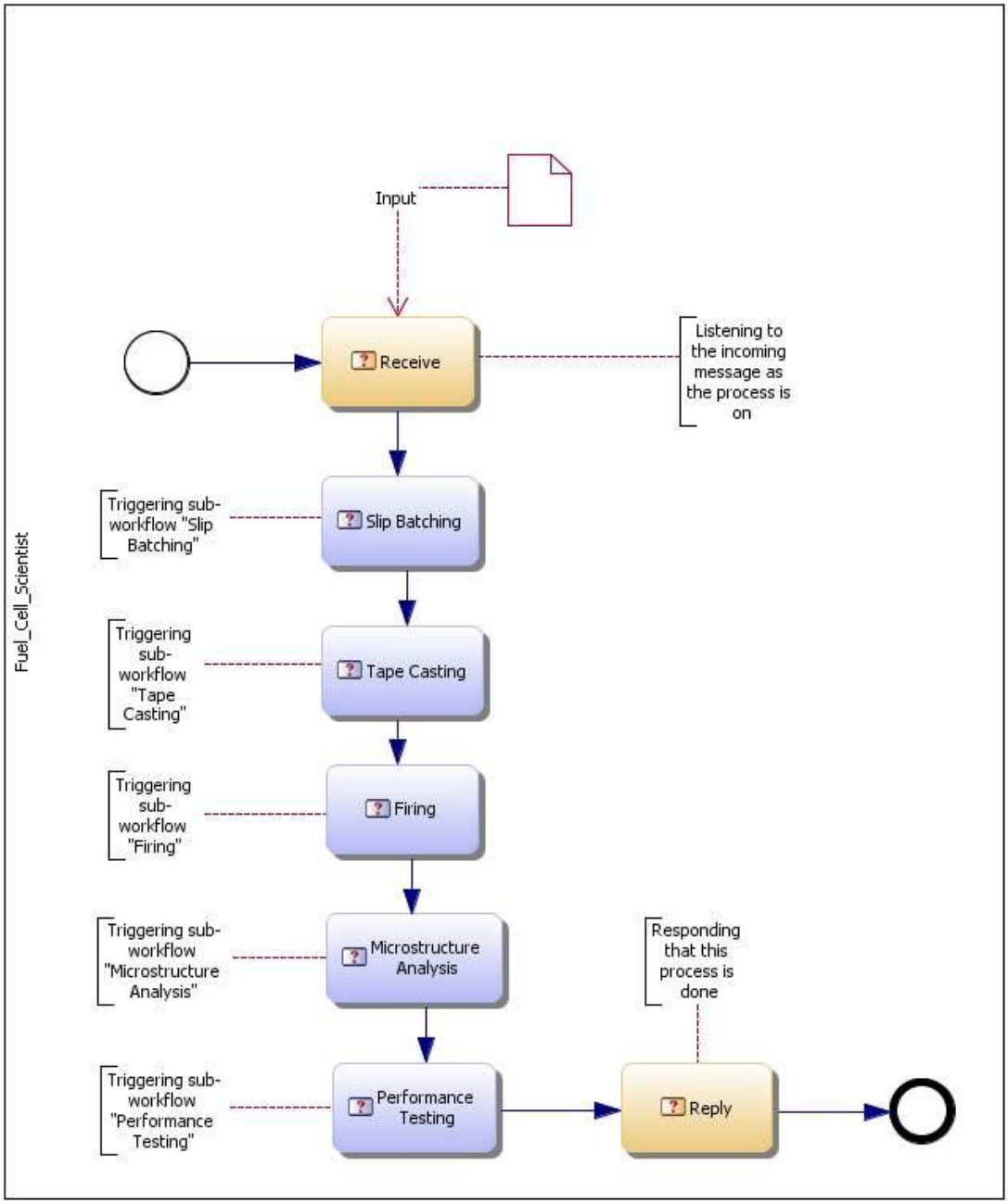


Figure B.1: A High Level View of Compound Synthesis Program

Figure A.2 demonstrates the low level view of the tape-casting sub-process, one of the human operation activities within the compound synthesis workflow. The others include the slip-batching, firing and performance testing activities.

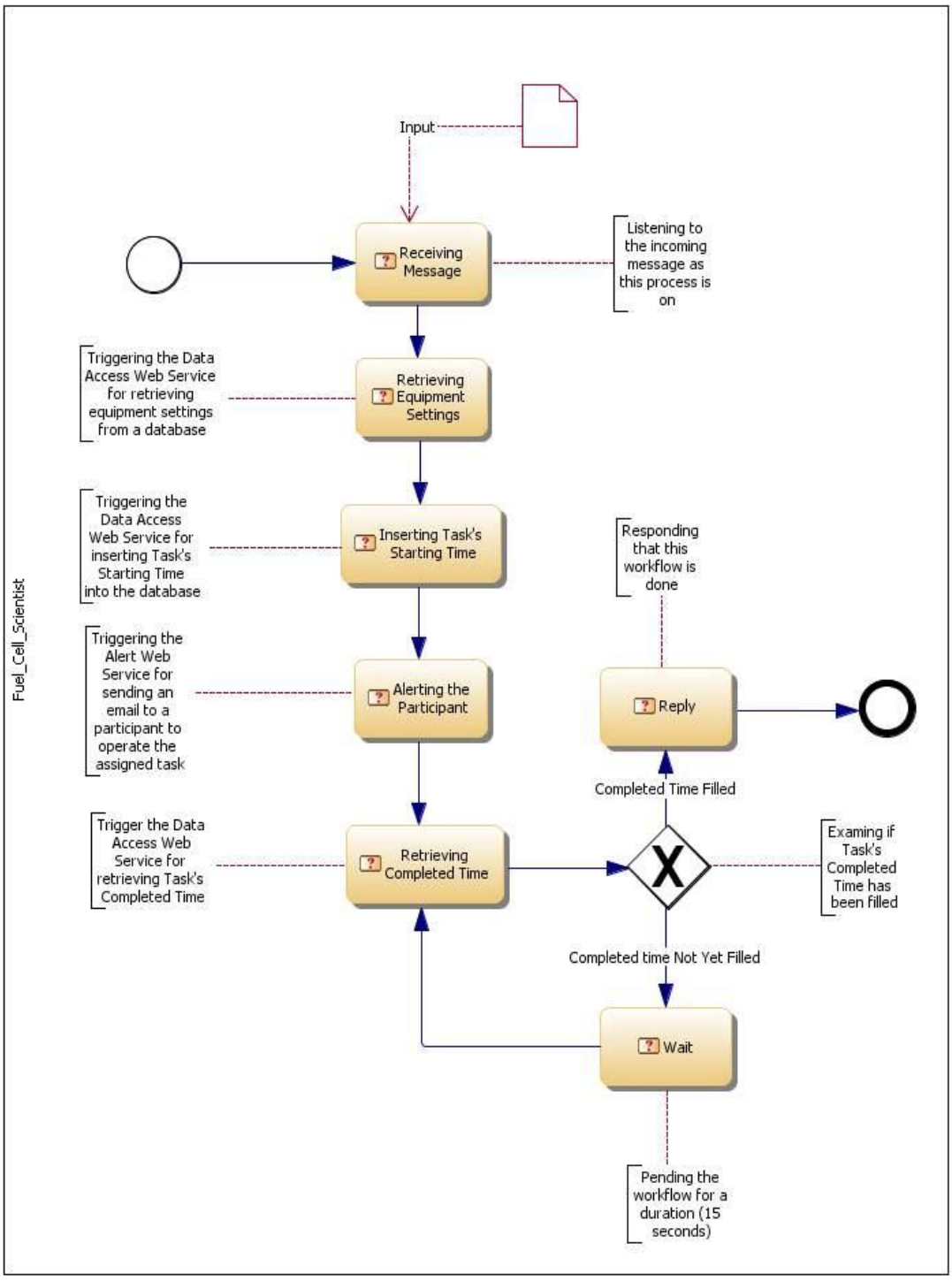


Figure B.2: Human-activity Sub-process – Tape Casting

Figure A.3 demonstrates the computational-activity sub-process.

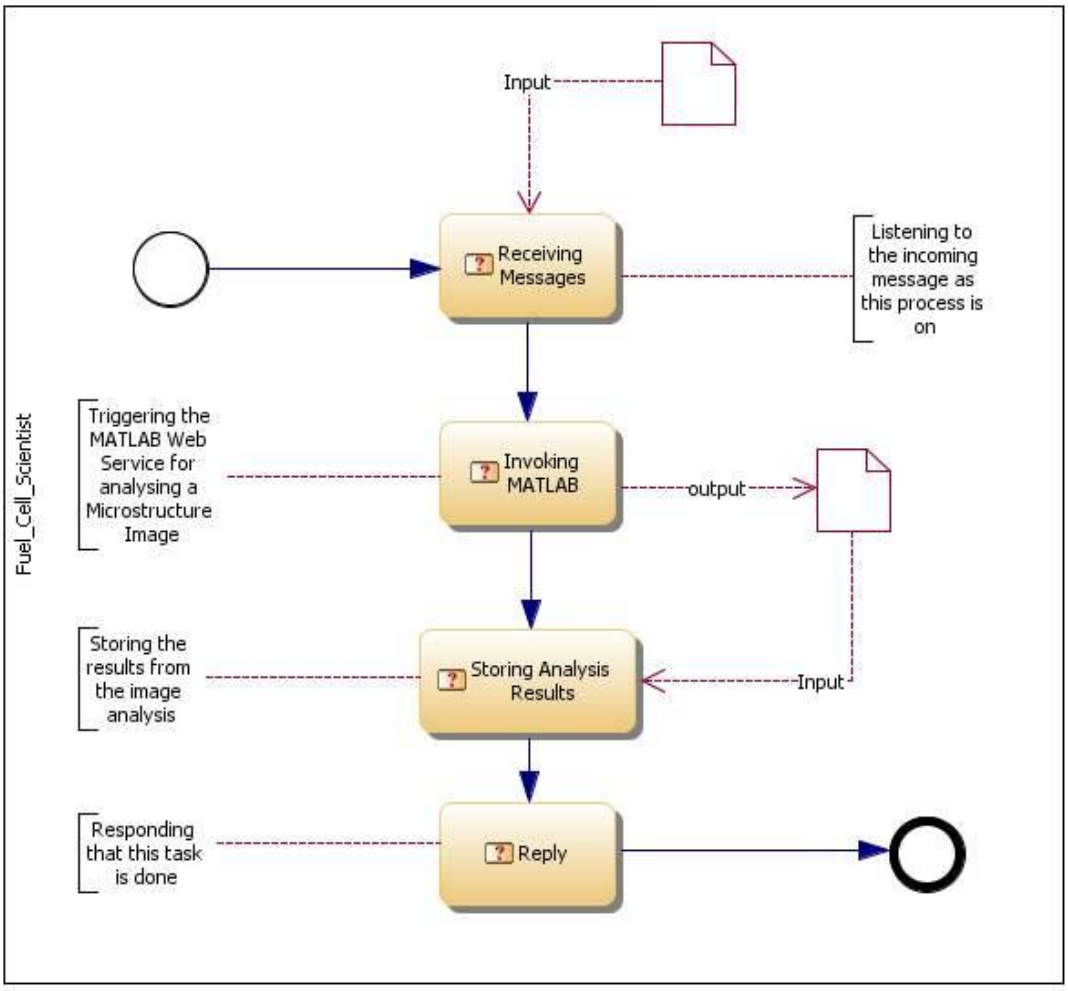


Figure B.3: Microstructure Analysis

Appendix C: D2R MAP Example File

```
<?xml version="1.0"?>
<d2r:Map xmlns:d2r="http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RMap/0.1#"
        d2r:versionInfo="$Id: eShopDB.Map.d2r, v 1.0 2003/01/20 19:44:09 Chris Exp $">

        <d2r:DBConnection   d2r:jdbcDriver="com.mysql.jdbc.Driver"
        d2r:jdbcDSN="jdbc:mysql://localhost:3306/compound_synthesis"
                                d2r:username="root"
                                d2r:password="" />

<d2r:Namespace d2r:prefix="comp_syn"
d2r:namespace=
"http://www.aibn.uq.edu.au/cmm/oxide_compound/synthesis.owl#"/>
        <d2r:Namespace d2r:prefix="foaf"
                        d2r:namespace="http://xmlns.com/foaf/0.1/#"/>
        <d2r:Namespace d2r:prefix="rdf"
                d2r:namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
        <d2r:ClassMap d2r:type="comp_syn:powder"
                d2r:sql="SELECT mixing.sample_id, mixing.batch_num, powder
                        FROM mixing, mixing_powder
                        WHERE mixing.sample_id=mixing_powder.sample_id AND
                                mixing.sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
                <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                        d2r:column="sample_id"/>
                <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                d2r:column="batch_num"/>
                <d2r:DatatypePropertyBridge d2r:property="comp_syn:powderName"
                                d2r:column="powder"/>
```

```
        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:input_to_synthesis"
                                      d2r:referredClass="comp_syn:synthesis"
        d2r:referredGroupBy="sample_id"/>

</d2r:ClassMap>


<d2r:ClassMap d2r:type="comp_syn:synthesis"

        d2r:sql="SELECT sample_id, batch_num

                FROM experiment

                WHERE sample_id='Melox2A-1475-4h'"

                        d2r:groupBy="sample_id">

        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"

                                      d2r:column="sample_id"/>

        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"

                                      d2r:column="batch_num"/>

        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:composed_of_SB"
                                      d2r:referredClass="comp_syn:slipBatching"

                                      d2r:referredGroupBy="sample_id"/>

        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:composed_of_TC"
                                      d2r:referredClass="comp_syn:tapeCasting"
                d2r:referredGroupBy="sample_id"/>

        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:composed_of_F"

                                      d2r:referredClass="comp_syn:firingActivity"

                                      d2r:referredGroupBy="sample_id"/>

        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:outputs_oxide"

                                      d2r:referredClass="comp_syn:compoundOutput"

        d2r:referredGroupBy="sample_id"/>

    </d2r:ClassMap>


    <d2r:ClassMap          d2r:type="comp_syn:slipBatching"

                d2r:sql="SELECT sample_id, batch_num

FROM slip_batching WHERE sample_id='Melox2A-1475-4h'"

                d2r:groupBy="sample_id">
```

```
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                        d2r:column="sample_id"/>
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                        d2r:column="batch_num"/>
            <d2r:ObjectPropertyBridge     d2r:property="comp_syn:SB_follows_TC"
                                        d2r:referredClass="comp_syn:tapeCasting"
                                        d2r:referredGroupBy="sample_id"/>
        </d2r:ClassMap>

        <d2r:ClassMap          d2r:type="comp_syn:tapeCasting"
                    d2r:sql="SELECT sample_id, batch_num
                            FROM tape_casting
                            WHERE sample_id='Melox2A-1475-4h'"
                    d2r:groupBy="sample_id">
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                        d2r:column="sample_id"/>
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                        d2r:column="batch_num"/>
<d2r:ObjectPropertyBridge     d2r:property="comp_syn:TC_follows_F"
d2r:referredClass="comp_syn:firingActivity"
d2r:referredGroupBy="sample_id"/>
        </d2r:ClassMap>

        <d2r:ClassMap          d2r:type="comp_syn:firingActivity"
                    d2r:sql="SELECT sample_id, batch_num
                             FROM firing
                            WHERE sample_id='Melox2A-1475-4h'"
            d2r:groupBy="sample_id">
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                        d2r:column="sample_id"/>
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
```

```
                    d2r:column="batch_num"/>
</d2r:ClassMap>

<d2r:ClassMap       d2r:type="comp_syn:compoundOutput"
            d2r:sql="SELECT sample_id, batch_num
                    FROM experiment
                    WHERE sample_id='Melox2A-1475-4h'"
            d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                    d2r:column="batch_num"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:input_to_characterization"
                    d2r:referredClass="comp_syn:characterization"
                    d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>
<d2r:ClassMap       d2r:type="comp_syn:characterization"
            d2r:sql="SELECT sample_id, batch_num
                    FROM experiment
                    WHERE sample_id='Melox2A-1475-4h'"
            d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                    d2r:column="batch_num"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:composed_of_TE"
                    d2r:referredClass="comp_syn:thermalExpansion"
                    d2r:referredGroupBy="sample_id"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:composed_of_EC"
                    d2r:referredClass="comp_syn:electronicConductivity"
```

```
                    d2r:referredGroupBy="sample_id"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:composed_of_EM"
                    d2r:referredClass="comp_syn:electronicMicroscopy"
                    d2r:referredGroupBy="sample_id"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:composed_of_XRay"
                    d2r:referredClass="comp_syn:X-RayDiffraction"
                    d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>

<d2r:ClassMap       d2r:type="comp_syn:X-RayDiffraction"
            d2r:sql="SELECT sample_id, batch_num
                    FROM xray_diffraction
                    WHERE sample_id='Melox2A-1475-4h'"
            d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                    d2r:column="batch_num"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:outputs_cell_pattern"
                    d2r:referredClass="comp_syn:cellPattern"
                    d2r:referredGroupBy="sample_id"/>
        <d2r:ObjectPropertyBridge   d2r:property="comp_syn:outputs_cell_param"
                    d2r:referredClass="comp_syn:cellParameter"
                    d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>

<d2r:ClassMap       d2r:type="comp_syn:cellPattern"
            d2r:sql="SELECT sample_id, batch_num, xrd_pattern
                    FROM xray_diffraction
                    WHERE sample_id='Melox2A-1475-4h'"
```

```
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                    d2r:column="batch_num"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:patternURI"
                                    d2r:column="xrd_pattern"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:cellParameter"

                d2r:sql="SELECT sample_id, batch_num, unit_cell_parameter
                         FROM xray_diffraction
                        WHERE sample_id='Melox2A-1475-4h'"
                                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                    d2r:column="batch_num"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:paramURI"
                                    d2r:column="unit_cell_parameter"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:cellDiagram"
                d2r:sql="SELECT sample_id, batch_num, unit_cell_diagram
                        FROM xray_diffraction
                        WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                    d2r:column="batch_num"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:diagramURI"
                                    d2r:column="unit_cell_diagram"/>
        <d2r:ObjectPropertyBridge    d2r:property="comp_syn:cell_derived_from"
                                   d2r:referredClass="comp_syn:cellParameter"
                                   d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:thermalExpansion"
                d2r:sql="SELECT sample_id, batch_num
                        FROM thermal_expansion
                        WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
```

```
                                    d2r:column="batch_num"/>
        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:outputs_TE_C"
                                    d2r:referredClass="comp_syn:TECoefficientDiagram"
                                    d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:TECoefficientDiagram"
                d2r:sql="SELECT sample_id, coefficient_diagram
                        FROM thermal_expansion
                        WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:diagramURI"
                                    d2r:column="coefficient_diagram"/>
</d2r:ClassMap>


<d2r:ClassMap          d2r:type="comp_syn:electronicConductivity"
                d2r:sql="SELECT sample_id, batch_num
                        FROM electronic_conductivity
                         WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                    d2r:column="batch_num"/>
        <d2r:ObjectPropertyBridge     d2r:property="comp_syn:outputs_EC"
                                    d2r:referredClass="comp_syn:conductivityNum"
                                   d2r:referredGroupBy="sample_id"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:conductivityNum"
                d2r:sql="SELECT sample_id, conductivity_data
                        ROM electronic_conductivity
                        WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:numURI"
                                    d2r:column="conductivity_data"/>
</d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:conductivityDiagram"
                d2r:sql="SELECT sample_id, conductivity_diagram
                        FROM electronic_conductivity
                        WHERE sample_id='Melox2A-1475-4h'"
                d2r:groupBy="sample_id">
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                    d2r:column="sample_id"/>
        <d2r:DatatypePropertyBridge d2r:property="comp_syn:condDiagramURI"
                                     d2r:column="conductivityl_diagram"/>
```

```
            <d2r:ObjectPropertyBridge      d2r:property="comp_syn:EC_derived_from"
                                           d2r:referredClass="comp_syn:conductivityNum"
                                           d2r:referredGroupBy="sample_id"/>
        </d2r:ClassMap>

        <d2r:ClassMap          d2r:type="comp_syn:electronicMicroscopy"
                       d2r:sql="SELECT sample_id, batch_num
                                FROM electronic_microscopy
                                WHERE sample_id='Melox2A-1475-4h'"
                       d2r:groupBy="sample_id">
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                        d2r:column="sample_id"/>
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:batchNumber"
                                        d2r:column="batch_num"/>
            <d2r:ObjectPropertyBridge      d2r:property="comp_syn:outputs_HRTEM"
                                           d2r:referredClass="comp_syn:EMImage"
                                           d2r:referredGroupBy="sample_id"/>
        </d2r:ClassMap>

<d2r:ClassMap          d2r:type="comp_syn:EMImage"
                       d2r:sql="SELECT sample_id, HRTEM_image
                                FROM electronic_microscopy
                                WHERE sample_id='Melox2A-1475-4h'"
                       d2r:groupBy="sample_id">
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:sampleID"
                                       d2r:column="sample_id"/>
            <d2r:DatatypePropertyBridge d2r:property="comp_syn:imageURI"
                                       d2r:column="HRTEM_image"/>
        </d2r:ClassMap>

</d2r:Map>
```